Det Nationale Forskningscenter
for Arbejdsmiljø

Symposium i anvendt statistik 2019

Økonomisk Institut, Københavns Universitet
Det Nationale Forskningscenter for Arbejdsmiljø

SYMPOSIUM I ANVENDT STATISTIK

28.-30. januar 2019

Redigeret af Peter Linde
på vegne af organisationskomiteen

Støttet af SAS Institute Inc.

Økonomisk Institut, Københavns Universitet
og
Det Nationale Forskningscenter for Arbejdsmiljø

Forord

Det er symposiets formål at fremme information om såvel anvendt statistik som statistisk databehandling. Symposiet er tværfagligt med særlig vægt på metodik, formidling og fortolkning af statistiske analyser. I år er Økonomisk Institut, Københavns Universitet vært for symposiet, hvilket vi gerne vil takke for. Symposiet arrangeres af Det Nationale Forskningscenter for Arbejdsmiljø, og Økonomisk Institut, Københavns Universitet. Den faglige forening Symposium i Anvendt Statistik er ansvarlig for det faglige program og økonomien.

Denne publikation indeholder foredragene fra det 41. Symposium i Anvendt Statistik. Dette års indlæg kommer fra mange forskellige fagområder og lægger vægt på forskellig metoder og problemstillinger. Som det er normalt ved videnskabelige indlæg, er bidragsyderne ansvarlige for indholdet af indlæggene, og spørgsmål herom kan rettes direkte til forfatterne.

Med symposiet tilstræbes det at skabe et forum for tværfaglig inspiration og kritik blandt andet for at udbygge kommunikationen mellem personer, der arbejder med beslægtede metoder inden for forskellige fagområder.

Peter Linde, Organisationskomiteen

ISBN 978-87-7904-359-6

Trykt hos PRinfoTrekroner i 200 eksemplarer

Organisationskomiteen for Symposium i Anvendt Statistik 2019

Lisbeth la Cour
Økonomisk Institut
Copenhagen Business School
Porcelænshaven 16A
2000 Frederiksberg
llc.eco@cbs.dk

Peter Linde
Det Nationale Forskningscenter
for Arbejdsmiljø
Lersø Parkallé 105
2100 København Ø
pli@nfa.dk

Anders Milhøj
Økonomisk Institut
Københavns Universitet
Øster Farimagsgade 5, bygning 26
1353 København K
Anders.Milhoj@econ.ku.dk

Esben Høg
Institut for Matematiske Fag
Aalborg Universitet
Skjernvej 4A
9220 Aalborg Ø
esben@math.aau.dk

Gorm Gabrielsen
Institut for Finansiering
Copenhagen Business School
Solbjerg Plads 3
2000 Frederiksberg
stgg@cbs.dk

Anders Holm
Sociologisk Institut
København Universitet
Øster Farimagsgade 5
1353 København K
ah@soc.ku.dk

Helle M. Sommer
SEGES
Landbrug & Fødevarer
Axeltorv
1609 København V
hmso@seges.dk

Niels Kærgaard
Fødevare- og Ressourceøkonomi
Københavns Universitet
Rolighedsvej 25
1958 Frederiksberg
nik@life.ku.dk

Mogens Dilling-Hansen
Institut for Økonomi
Århus Universitet
8000 Århus C
dilling@econ.au.dk

Klaus Rostgaard
Statens Serum Institut
Artillerivej 5
2100 København Ø
KLP@ssi.dk

Jørgen Lauridsen
Økonomisk Institut
Syddansk Universitet
Campusvej 55
5230 Odense M
jtl@sam.sdu.dk

Kristina Birch
SAS Institute
Købmagergade 7-9
1050 København K
Kristina.Birch@sdk.sas.com

Birthe Lykke Thomsen
Ferring
Kay Fiskers Plads 11
2300 København S
Birthe.Thomsen@ferring.com

Indholdsfortegnelse

Exploratory methods, ML and prediction

Sundhed, socialområdet og SAS

Uddannelse

Anvendt statistik

Økonomi og forbrug

ON OMITTED VARIABLES, PROXIES AND UNOBSERVED EFFECTS IN ANALYSIS OF ADMINISTRATIVE LABOUR MARKET DATA

Author: Shihan Du; Pia Homrighausen, Ralf A. Wilke

Abstract: In virtually any empirical regression analysis, there is limited availability of observed variables and limited prior knowledge on which variables belong to the model. This paper provides a unified framework that nests various approaches aiming at reducing omitted variable bias in linear regression analysis. We work out the mechanisms driving the size of the bias and how various models with different regressor sets or unobserved effects relate. Without imposing restrictions on the relationship and role of the variables, it is, however, not possible to derive model rankings that are valid in every application.

In our applications, we apply the various models to German linked administrative labour market data. We find evidence for sizeable omitted variable bias in a wage regression, while a labour market transition model appears to be less affected. Additional survey variables contribute only to the wage model, while the use of work history variables and panel models lead to changes in coefficients in the two models. Overall, panel data models with a restricted regressor set are found to control for more unobserved information than cross-sectional analysis with an extended variable set.

Relative Importance

Christine Winther Bang[1], Volkert Siersma[1]

[1]The Research Unit for General Practice and Section of General Practice, Department of Public Health, University of Copenhagen, Copenhagen, Denmark

Often the primary target of an analysis is to know which of a collection of factors are most important for a certain quality indicator. If these factors can be influenced, such assessment gives leads as to which interventions may give the largest improvement in the quality indicator. The traditional approach is to use repeated hypothesis testing in some model search procedure (see Bring 1996 for a discussion) and thereby analyze something close to – but subtly different from – the target of the analysis. With this traditional approach, we may find a set of factors that predict the quality indicator well, but, notably when factors are correlated among themselves, other sets may do just as well and factors dropped in the process may well be important and valuable as targets of intervention.

A way to take the target of analysis head on is to assess a relative importance of the factors. For a continuously valued quality indicator the relative importance of factors is the decomposition of the coefficient of determination (R^2) from a multivariable linear regression model according to the contribution of the factors. In its most basic form, this decomposition is calculated by averaging the increase in R^2 values from adding a factor over all possible multivariable linear regression models. This approach has been explored and implemented by others for the linear regression model.

The assessment of relative importance comprises an exhaustive walkthrough of all possible models, which can get computationally heavy if a high number of factors is investigated. Hence, in the present report we explore Monte Carlo methods to assess relative importance. Furthermore, through the use of various pseudo R^2 coefficients the relative importance assessment is readily expanded to generalized linear models, in which a Monte Carlo implementation proves invaluable.

PORTAAT

The method to assess relative importance, and various issues with and improvements to this method, are illustrated with 2015 data from the PORTAAT study. This study was conducted among employees of the city of Pori, in South-Western Finland in the years 2014 and 2015. In 2015, 671 workers (76 men and 595 women, median age 50 [IQR 42-56, range 20-68]) from ten work units participated; occupations included librarians, museum employees, ground

keepers, computer workers, social workers, nurses, physicians, administrative officials and general office staff. Data was collected in a physical examination and with questionnaires. See Veromaa et al. (2017a, 2017b, 2017c) for details.

The aim in the present investigation is to evaluate the importance of various work-related, social, life-style and health-related factors on general life satisfaction. The motivation for this is that workers who are satisfied, not only with the work they are doing, but with the life situation in general, will be advantageous for the employer. If we know the drivers of life satisfaction, organizational change can target these drivers so as to improve life satisfaction for the Pori municipal workers.

General life satisfaction is measured by the EuroHIS (Schmidt et al 2006) scale. This scale, range 1=lowest satisfaction to 5=highest satisfaction, is calculated as the mean of eight items on satisfaction with various aspects of life (e.g. health, relationships, the neighborhood) each scoring 1-5 also. The median EuroHIS in the 2015 PORTAAT data was 4.13 (IQR 3.75-4.38, range 2.25-5.00). Alongside the EuroHIS scale, a EuroHIS dissatisfaction indicator is defined by scoring 1 or 2 in at least one of the eight EuroHIS items. With this definition 136 (20.3%) employees indicated dissatisfaction.

The various drivers of life satisfaction that are evaluated in the following are divided into roughly four themes:

1) Work-related factors: *Burnout* measured by the Bergen Burnout Indicator (Salmela-Aro et al 2011), *Work engagement* measured by the Utrecht Work Engagement Scale (Schaufeli et al 2006) implemented both as a total score and as three sub-scores: Vigor, Dedication and Absorption, and *Daytime work* indicating whether the work is performed in normal office hours.
2) Social factors: *Social support* measured by the mean of the six ENRICHD items (ENRICHD 2001), *Cohabiting* indicating whether the employee lives with others, and *Enough money* indicating content with the employee's financial situation.
3) Life-style factors: *Alcohol* measured with AUDIT (Aalto et al 2006) or with a dichotomization of the AUDIT scale indicating risky alcohol habits, *Smoking* indicating current, former and never smokers, Body-Mass Index (*BMI*) calculated as body weight (in kg) divided by height (in m) squared, and Resting Heart Rate (*RHR*) in bpm.
4) Health-related factors: *Absenteeism* calculated as the number of days absent from work in the two year period 2014-2015, *Often headache* measured by the first item of the Headache Impact Test (Kosinski et al 2003), *Medication*

indicating regular use of prescription drugs, and *Multimorbidity* defined as the number of chronic diseases.

Apart from the above potential drivers, sex, education and age are included in the regression models, but are not evaluated on their importance for life satisfaction.

LMG

The LMG method (Lindeman 1980) is the method to assess relative importance that is consistently recommended (e.g. Johnson & LeBreton 2004, Grömping 2006). This method calculates for each of the factors that are to be assessed the increase in R^2 when adding this factor to each of the possible multivariable models composed of subsets of the other factors that are assessed. These increases are then averaged over the factors, giving a decomposition of the full model R^2. The relative importance is then given as a percentage of the full model R^2. This setup is surprisingly elegant and simple to implement. Furthermore, it is straightforwardly outfitted with useful features, e.g. various factors can be grouped so that their joint importance is assessed and not the importance of the individual factors, and certain factors can be included in all regression models and taken out of the relative importance assessment (Grömping 2006). The latter is the case for sex, education and age in the present investigation: while these are probably important for life satisfaction, they are not easily targeted in organizational change in any productive way. Hence, the relative importance of the potential drivers for life satisfaction is based on a decomposition of the full model R^2 minus the R^2 of a model with only sex, education and age.

The LMG method needs an exhaustive walkthrough of all possible models. With n factors to be evaluated, there are

$$(1) \quad \binom{n-1}{k} = \frac{(n-1)!}{k!(n-k-1)!}$$

models with k, $0 \leq k < n$, factors to be evaluated for the increase in R^2 when another factor is added. Hence, for a selection of n factors, an exhaustive walkthrough needs

$$(2) \quad \sum_{k=0}^{n-1} \binom{n-1}{k} n = \sum_{k=0}^{n-1} \frac{n!}{k!(n-k-1)!} = \sum_{k=0}^{n-1} \binom{n}{k} (n-k)$$

evaluations; this amounts to 192 for six factors, and 524,288 for 16 factors. Multivariable linear regression is quite fast, but in practice, the 20 factors evaluated (10,485,760 evaluations) in Siersma (2013) took 10 minutes for the Grömping (2006) implementation of LMG; an additional factor would increase this time prohibitively.

Monte Carlo

A Monte Carlo implementation of the LMG method alleviates the computational burden of the exhaustive walkthrough implementation when the number of factors to be evaluated is high. The LMG method calculates averages of R^2 increases over all possible multivariable models. Instead of calculating all R^2 increases and taking their average, these averages can be estimated in a (random) subset of all possible multivariable models. These subsets may imply considerable fewer evaluations than the exhaustive walkthrough. Furthermore, these averages can be calculated incrementally so as to be able to stop when estimates do not change anymore.

A Monte Carlo version of the LMG method is easily implemented, but details are not entirely trivial. First, the random sampling needs to be done so that for a certain factor (out of n possible factors), the model that is sampled is of size k with probability

$$(3) \quad \frac{\binom{n-1}{k}}{\sum_{l=0}^{n-1}\binom{n-1}{l}n}$$

which is found by combining formulae (1) and (2). This weighs down the smaller and larger models. If, naively, each size of model is sampled with equal probability, both small and large models are selected too often, which, in the best case, will draw out the simulations unnecessarily long and, in the worst case, will give wrong relative importance. Second, it is not trivial when to stop the simulation. Intuitively, the simulation can stop when subsequent iterations do not result in much change in the relative importance assessment anymore. For an iteration $i > 0$ we stop the simulation when

$$(4) \quad \frac{\sum_{l=1}^{n}\left(\frac{\sum_{j=1}^{i}\left(\Delta R^2_{j-1,l}-\Delta R^2_{j,l}\right)}{i}\right)^2}{\sum_{l=1}^{n}\left(\frac{\sum_{j=1}^{i}\Delta R^2_{j,l}}{i}\right)^2} < \varepsilon$$

where $\Delta R^2_{j,l}$ is the increase in R^2 when adding factor l in the jth iteration, $\Delta R^2_{j,l} = 0$ for $j = 0$, and ε is a tolerance level which was set to 0.00001. While this criterion seems to function well in the evaluations below, it remains to be seen whether this stopping criterion is the best choice in general.

Small example

We used the LMG method on a selection of six factors, the work-related and most of the social factors, from the 2015 PORTAAT data to assess their relative

importance to general life satisfaction. Sex, age and education are also in the model, but are not evaluated for relative importance. The full model has an R^2 of 0.3515 and the model with only sex, age and education has an R^2 of 0.0098. Hence, the relative importance of the six factors is based on a decomposition of $\Delta R^2 = 0.3417 = 0.3515 - 0.0098$, the proportion of variance explained by the six factors. The results for the exhaustive walkthrough implementation are shown in the first column of Table 1. This took 11 seconds to compute.

Table 1: Relative importance of selected work-related and social factors for life satisfaction.

	Exhaustive walkthrough	Monte Carlo
Burnout	37.1%	39.1%
Social support	29.2%	27.2%
Work engagement	26.6%	27.0%
Absenteeism	5.7%	5.4%
Cohabiting	1.1%	1.0%
Daytime work	0.3%	0.3%

The second column in Table 1 shows the results computed by the Monte Carlo implementation. This computation took 390 iterations, but the Monte Carlo implementation in general may need more than 1000 iterations to reach the stopping criterion; with realistic computation times of 27 seconds. The convergence is shown in Figure 1.

Figure 1: Convergence of the Monte Carlo estimates of relative importance of selected work-related and social factors for life satisfaction.

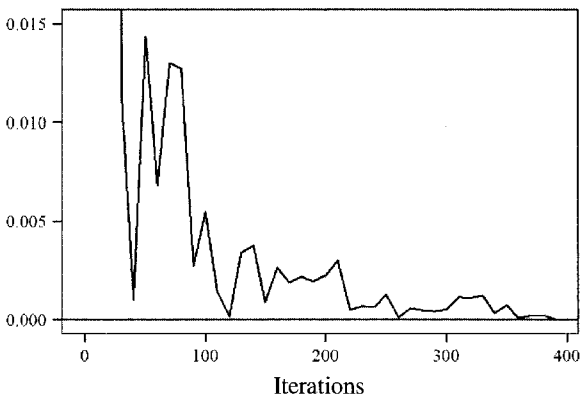
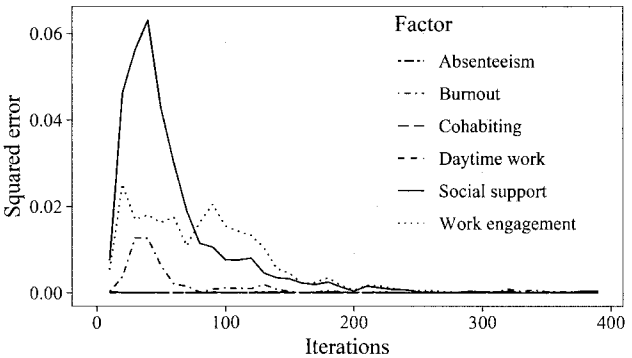


Figure 2: Convergence of the Monte Carlo estimates of relative importance of selected work-related and social factors for life satisfaction relative to the corresponding estimates from an exhaustive walkthrough.



While the differences in relative importance assessments seen in Table 1 are small, the improvement in computation time by using the Monte Carlo assessment is doubtful. However, the purpose of the present analysis is to externally evaluate the convergence of the Monte Carlo implementation of LMG. For this limited selection of six factors the exhaustive walkthrough implementation is readily computed and the difference with the Monte Carlo implementation can be evaluated. In Figure 2 the convergence of the Monte Carlo estimates to the exhaustive walkthrough estimates is shown for each of the six factors separately as the squared error relative to the exhaustive walkthrough assessment in the average increase in R^2 over the iterations. Convergence is seen adequate, and the Monte Carlo estimates are already close to the estimates from the exhaustive walkthrough after 200 iterations. The stopping criterion (4) functions well in this case.

Relative importance features

The Monte Carlo implementation of the LMG method can be used to assess the relative importance for a larger selection of factors so as to elevate the previous toy example to a more useful assessment. In Table 2 the relative importance of the previously presented work-related, social, life-style and health-related factors for general life satisfaction is assessed. For the most basic configuration of factors this is shown in the column marked Model 1. This set of 14 factors needed 114,688 evaluations, taking more than an hour, in an exhaustive walkthrough implementation, but the Monte Carlo evaluation converged in 1710 iterations taking 71 seconds; a considerable improvement. The conclusion from

this evaluation could be that in order to improve life satisfaction among municipal workers in Pori, interventions affecting Burnout, Work engagement, Social support and Enough money are suggested, while interventions on life-style factors such as smoking and risky alcohol habits will not affect life satisfaction much.

Table 2: Relative importance of work-related, social, life-style and health-related factors for life satisfaction.

	Model 1	Model 2		Model 3
Burnout	25.0%	20.7%		20.7%
Work engagement	19.4%			
Work engagement (Vigor)		17.4%		11.9%
Work engagement (Dedication)		7.1%	26.8%	7.7%
Work engagement (Absorption)		2.3%		1.4%
Daytime work	0.1%	0.1%		0.0%
Social support	19.7%	20.4%		22.0%
Cohabiting	0.5%	0.6%		0.5%
Enough money	16.6%	13.3%		15.6%
Alcohol	0.1%	0.1%		
AUDIT				2.2%
Smoking	0.6%	0.7%		0.6%
BMI	3.8%	4.6%		4.3%
RHR	0.1%	0.1%		0.2%
Absenteeism	1.9%	2.2%		1.7%
Often headache	2.2%	2.2%		2.1%
Medication	0.9%	0.8%		0.8%
Multimorbidity	9.1%	7.2%		8.2%

The manner in which the factors affect life satisfaction is not read from the relative importance assessment and has to be obtained by investigating an appropriate regression, e.g. a regression including only the most important factors. This is rather important because the relative importance may cover over some counterintuitive associations. For example, smokers may rather enjoy their small breaks and nicotine rushes that add to their life satisfaction so that an intervention to reduce smoking may have an effect opposite as intended. In the PORTAAT 2015 data this seems not the case: the factor Smoking was not very important and we found in a linear regression that smokers had lower life satisfaction.

The relative importance among a set of factors is a characteristic of the data for which it was assessed. For an individual, a certain factor, e.g. a diagnosis of

serious disease, may affect life satisfaction a lot, but if this factor does not appear much in our data, or not at all, the factor is still not assessed important in our data. Intuitively, improvement in such factor would increase life satisfaction for only few people and not increase the average life satisfaction for the group much. This underlines the need for the data to be representative of some well-defined group of people. In this sense the PORTAAT data is a representative sample of employees of Pori municipality in 2015 and therefore can be used to find areas of organizational change to the largest effect on employees' life satisfaction.

Relative importance can be boosted by giving a factor more degrees of freedom. Hence, this should be done with caution. A guiding principle is that a factor should be featured in an assessment in a manner that an organizational change can affect it. For example, Work engagement may be divided into three different aspects. In Table 2 in the column marked Model 2 we can see that the combined importance relative to the other factors rises to 17.4+7.1+2.3=26.8%; this is partly on the account of the increased degrees of freedom for the three combined Work engagement factors relative to the single one in Model 1. However, we learn that Vigor is more important than Dedication or Absorption, and this is a useful lead for organizational change. In Table 2 in the column marked Model 3, the Alcohol factor is switched with AUDIT. The Alcohol factor is a dichotomization of AUDIT indicating risky alcohol use; AUDIT also categorizes low risk alcohol use. As expected because of the increased degrees of freedom, the relative importance of AUDIT is higher than of Alcohol. The organizational change envisaged is different: changing risky alcohol use or changing alcohol use in general? The former is tentatively the more obtainable goal. Note also that no alcohol use may well be related to illness.

Generalized linear models

The assessment of relative importance is straightforwardly adapted to generalized linear models, except for the fact that there is no single coefficient of determination. Generalized linear models, for example logistic regression or Poisson regression, are useful if the assumptions of the linear regression model are not met and the outcome is better modeled with a link function other than the identity link. Generalized linear models often take somewhat longer to estimate so that we can put the Monte Carlo implementation to good use. For generalized linear models, instead of using the linear regression coefficient of determination, one can choose between quite some pseudo-R^2 coefficients. For logistic regression, the generalized linear regression we will focus on in the following, several measures of predictive ability reviewed in Mittlbock & Schemper (1996) and Menard (2000).

One pseudo-R^2 coefficient is attributed to McFadden (McFadden 1974). This coefficient penalizes a perfect fit with the quotient of the natural logarithm of the likelihoods of the full model and the empty model as in (5).

$$(5) \quad \text{McFadden } R^2 = 1 - \frac{\log(L_{model})}{\log(L_{empty})}$$

However, this pseudo-R^2 does not revert to the conventional R^2 as a special case if the model is the conventional linear regression model. This is achieved by the pseudo-R2 attributed to Cox and Snell (Cox & Snell 1989) which adjusts the penalty in (5) by the number of observations n as in (6).

$$(6) \quad \text{Cox} - \text{Snell } R^2 = 1 - \left(\frac{\log(L_{empty})}{\log(L_{model})}\right)^{2/n}$$

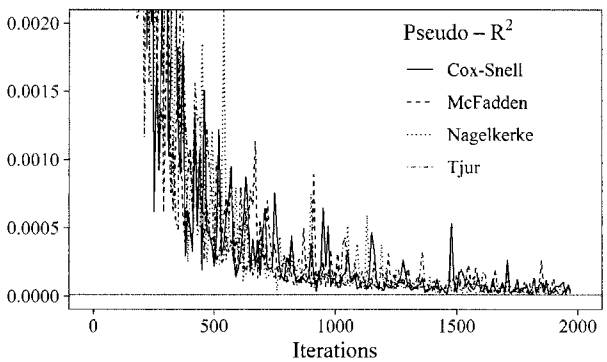
Hence, this pseudo-R^2 can be used for many link functions and can therefore be viewed as a generalized R^2 coefficient. The disadvantage with this pseudo-R2 is that its upper bound is lower than 1. This is corrected in a third pseudo-R^2 attributed to Nagelkerke (Nagelkerke 1991) which is just the Cox-Snell R^2 (6) normalized to have an upper bound of one. A fourth pseudo-R^2, specifically made to the logistic regression, is attributed to Tjur (Tjur 2009) and evaluates the estimated probabilities with the observed outcomes. In fact, this is the conventional R^2 from a linear regression where the binary outcome is treated as continuously valued.

In Table 3 the relative importance of the work-related, social, life-style and health-related factors for the EuroHIS dissatisfaction indicator is assessed using the four pseudo-R^2 coefficients presented above. The assessments with Cox-Snell R^2 and Nagelkerke R^2 are very similar, which is to be expected as the latter is just a normalized version of the former. Between the other pseudo-R^2 coefficients, we find the results are somewhat different, but the orderings of the factors are similar. Note that these assessments of relative importance are not expected to show the same results as the different pseudo-R^2 measure slightly different aspects of model fit. It is not obvious which pseudo-R^2 should be used and it may be useful to consider multiple assessments using different pseudo-R^2 coefficients. Even though, overall conclusions do not differ for the present data. Convergence speed is also similar, see Figure 3.

Table 3: Relative importance of work-related, social, life-style and health-related factors for dissatisfaction, according to each of the four pseudo-R^2 coefficients.

	McFadden	Cox-Snell	Nagelkerke	Tjur
Enough money	20.12%	23.89%	22.14%	18.98%
Work engagement (Vigor)	16.09%	13.53%	13.75%	18.27%
Multimorbidity	15.94%	15.03%	16.42%	14.80%
Burnout	10.09%	8.77%	10.46%	10.05%
BMI	9.96%	8.09%	8.94%	10.64%
Social support	8.55%	10.02%	10.14%	8.87%
Absenteeism	6.64%	6.80%	5.64%	7.20%
Work engagement (Dedication)	4.33%	4.35%	4.14%	4.28%
Cohabiting	2.27%	2.18%	1.91%	1.94%
Medication	2.06%	2.61%	1.81%	2.05%
Often headache	1.39%	1.29%	1.39%	0.98%
Work engagement (Absorption)	1.23%	2.08%	1.63%	1.04%
RHR	0.64%	0.70%	0.91%	0.25%
Alcohol	0.35%	0.34%	0.36%	0.35%
Smoking	0.22%	0.21%	0.26%	0.26%
Daytime work	0.12%	0.11%	0.11%	0.03%

Figure 3: Convergence of the stopping criterion for the four pseudo-R^2 measures.



Discussion

The convergence speed of the Monte Carlo implementation can be improved in several ways. Firstly, faster glm procedures such as in the R packages speedglm or glmnet, or a C++ implementation can be used. Secondly, if the design matrix is sparse, we could use special features of the sparse matrix algebra. Thirdly, the

number of iterations could be lowered, which may be done by investigating alternative stopping criteria or sampling without replacement.

The assessment of relative importance by LMG still works when the number of factors that are evaluated (or rather the corresponding degrees of freedom) is larger than the number of observations in the data. For models with more degrees of freedom than observations $R^2 = 1$ and adding a factor to this model increases R^2 with 0. Hence, a relative importance assessment is well defined also in this case. However, the properties of such an assessment are unclear: can we capture the relative importance in data where more factors than observations have effect on the outcome?

Instead of a generalized linear model, we can often use a linear model R^2 and the conventional LMG implementation additional to the various pseudo-R^2 coefficients available, even though the model assumptions may be violated. For the logistic regression model this is done in the Tjur pseudo-R^2. This may not be problematic as the assessment of relative importance only uses the R^2, which does not rely on the model assumptions.

Conclusion

In the present report, we have adapted the assessment of relative importance to a Monte Carlo evaluation, which makes it possible to evaluate a much larger number of factors than in the exhaustive walkthrough evaluation. In addition, this method is straightforwardly adapted to generalized linear models and other regression models.

Literature

Aalto M, Tuunanen M, Sillanaukee P, Seppa K (2006) Effectiveness of structured questionnaires for screening heavy drinking in middle-aged women. *Alcoholism: Clinical and Experimental Research* 30: 1884-1888.

Baltes BB, Parker CP, Young LM, Huff JW, Altmann R (2004) The practical utility of importance measures in assessing the relative importance of work-related perceptions and organizational characteristics on work-related outcomes. *Organizational Research Methods* 7: 326-340.

Bring J (1996) A geometric approach to compare variables in a regression model. *The American Statistician* 50: 57–62.

Cox DR, Snell EJ (1989) *Analysis of Binary Data*. Second Edition. Chapman & Hall.

ENRICHD investigators (2001) Enhancing recovery in coronary heart disease (ENRICHD) study intervention: rationale and design. *Psychosomatic Medicine* 63: 747-755.

Grömping U (2006) Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, 1.

Johnson JW, LeBreton JM (2004) History and use of relative importance indices in organizational research. *Organizational Research Methods* 7: 238-257.

Kosinski M, Bayliss MS, Bjorner JB, Ware JE Jr, Garber WH, Batenhorst A, Cady R, Dahlöf CG, Dawson A, Tepper S (2003) A six-item short-form survey for measuring headache impact: the HIT-6. *Quality of Life Research* 12: 963-974.

LeBreton JM, Binning JF, Adorno AJ, Melcher KM (2004) Importance of personality and job-specific affect for predicting job attitudes and withdrawal behavior. *Organizational Research Methods* 7: 300-325.

Lindeman RH, Merenda PF, Gold RZ (1980) *Introduction to Bivariate and Multivariate Analysis.* Glenview IL: Scott, Foresman.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Pp. 105-142 in P. Zarembka (ed.), *Frontiers in Econometrics*. Academic Press.

Menard, S. (2000) Coefficients of determination for multiple logistic regression analysis. *The American Statistician* 54: 17-24.

Mittlbock, M. and M. Schemper (1996) Explained variation in logistic regression. *Statistics in Medicine* 15: 1987-1997.

Nagelkerke NJD (1991) A note on a general definition of the coefficient of determination. *Biometrika* 78: 691-692.

Salmela-Aro K, Rantanen J, Hyvonen K, Tilleman K, Feldt T (2011) Bergen Burnout Inventory: reliability and validity among Finnish and Estonian managers. *International Archives of Occupational and Environmental Health* 84: 635-645.

Schaufeli W, Bakker A, Salanova M (2006) The measurement of work engagement with a short questionnaire: a cross-national study. *Educational and Psychological Measurement* 66: 701-716.

Schmidt S, Muhlan H, Power M (2006) The EUROHIS-QOL 8-item index: psychometric results of a cross-cultural field study. *European Journal of Public Health* 16: 420-428.

Siersma V, Thorsen H, Holstein PE, Kars M, Apelqvist J, Jude EB, Piaggesi A, Bakker K, Edmonds M, Jirkovska A, Mauricio D, Ragnarson Tennvall G, Reike H, Spraul M, Uccioli L, Urbancic V, van Acker K, van Baal J, Schaper NC (2013) The importance of factors determining the low health-related quality of life in people presenting with a diabetic foot ulcer: the Eurodiale study. *Diabetic Medicine* 30: 1382-1387.

Tjur T (2009) Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician* 63: 366-372.

Veromaa V, Kautiainen H, Saxen U, Malmberg-Ceder K, Bergman E, Korhonen PE (2017a) Ideal cardiovascular health and psychosocial risk factors among Finnish female municipal workers. *Scandinavian Journal of Public Health* 45: 50-56.

Veromaa V, Kautiainen H, Korhonen PE (2017b) Physical and mental health factors associated with work engagement among Finnish female municipal employees: a cross-sectional study. *BMJ Open* 7: e017303.

Veromaa V, Kautiainen H, Juonala M, Rantanen A, Korhonen PE (2017c) Self-rated health as an indicator of ideal cardiovascular health among working-aged women. *Scandinavian Journal of Primary Health Care* 35: 322-328.

Behaviour changes associated with diarrhoea and respiratory diseases in Danish pre-weaned dairy calves

Leslie Foldager[1,2*] & Margit Bak Jensen[1]

[1] Department of Animal Science, Aarhus University, DK8830 Tjele, Denmark
[2] Bioinformatics Research Centre, Aarhus University, DK8000 Aarhus C, Denmark
* Email: leslie@anis.au.dk

Abstract

To investigate effects of diarrhoea and respiratory diseases, milk feeding behaviour and activity were collected automatically using computer controlled milk feeders and leg attached accelerometers for 169 Danish group-housed pre-weaned dairy calves (126 large breeds, mainly Holstein, and 43 Jersey). The calves were assigned to either a high milk allowance (8 l/d for large breeds; 6.4 l/d for Jersey) or a low milk allowance (6.4 and 4.8, respectively). Each pen contained both breeds and a mixture of male and female calves but only one type of milk allowance (high or low). In this presentation, we consider data transformed into daily measurements of number and duration of unrewarded visits to the milk feeder (visits without being entitled to milk), number of postural changes, lying duration, and the ratio between milk intake and allowance.

The hypotheses were that 1) diseased calves have a reduced appetite and 2) diseased calves have a lower activity. The indication of the first hypothesis may be expected to depend on the milk allowance: 1a) reduced milk intake for calves on high allowance and 1b) reduced number and duration of unrewarded visits for calves on low allowance. These may also be seen as indication for the second hypothesis but in addition, we expect to observe 2a) increased lying duration and 2b) reduced frequency of postural changes. Since respiratory diseases often also bring fever, the lowered activity may be more pronounced here than in the case of diarrhoea.

Most calves entered the pens from around day 10 after birth (median=10, range: 9-21) and data collection ended after 47 days but the calves stayed in the pen. Four calves died prematurely (after 5, 7, 12 and 19 days). The health status of each calf was measured by clinical examination Monday, Wednesday and Friday each week. In addition to the milk allowance, sex, breed (large/Jersey), age at insertion, and age of the oldest calf in the pen at time of insertion were included in the models. The two age variables were standardised to zero mean and unit variance. In total, 151 calves had at least one instance of diarrhoea (without concurrent respiratory disease or fever), 49 had at least one case of lung disease (without concurrent diarrhoea), 40 had concurrent diarrhoea and lung disease at least once, and 10 had at least on occasion of fever (without diarrhoea or lung disease). To decrease the number of possible disease states, we decided to pool lung

diseases, lung disease with concurrent diarrhoea, and fever. We will refer to this as *other illness* and 58 calves had at least one such illness period.

We intend to present results from two or three sets of statistical analyses. First, we investigate by competing-risk regressions if risk of disease is different between groups (breed, sex, milk allowance) and whether monitoring of milk feeding behaviour and activity can be used for early detection of disease risk. Changes in explanatory variables from one day to the next was calculated and time-lagged measurements one to 6 days back in time will be examined. In these time-to-event models, we currently only consider time from entry into the study/pen and until first day with the disease of interest or until right censoring at culling, day 47 or first day with the competing disease.

Secondly, courses of the automatic measurements will be investigated around disease periods by examining the trends up to 6 days before disease and 6 days during the period of disease/recovery. Here we consider only days with status "healthy" from the pre-disease period and only days with the current disease or "healthy" status from the disease/recovery period. To this end, we applied (generalised) linear mixed models with random effects of pen and calf nested in pen.

Thirdly, we are also consider using machine learning methods, e.g. random forests, to explore the possibility of early detection of illness.

All of this is work in progress and we do not currently have results ready to present in written form. Offhand though, it appears that a single day drop in duration of unrewarded visits may indicate an increased risk of being diagnosed with diarrhoea approximately three days later.

An exact likelihood ratio test for Ancestry-informative markers

Torben Tvedebrink, Department of Mathematical Sciences, Aalborg University

Introduction

Ancestry-informative markers (AIMs) are genetic markers that give information about the ancestry of individuals. By ancestry we refer to the genogeographic origin, which is the term used link certain genetic characteristics with a specific geographic region, culture or ethnicity. The term biogeographical ancestry is also used in the literature, however, the inference of the ancestry may then also depend on non-genetic features.

AIMs markers are of interest to forensic geneticists as they can provide investigative leads to the police regarding the ancestry of an unknown culprit or unidentified human remains. This may be valuable in crime cases where there is no other type of information available, e.g. no witness testimonies, to narrow down the group of potential suspects. It can also be applied in mass disaster victim identification cases as a first means of identifying the body remains. The prevailing genetic marker used in ancestry are Single Nucleotide Polymorphisms (SNPs), which are genetic markers with population variability on a single position. Most SNPs are bi-allelic, meaning that there are two variants. Hence, it is sufficient to count the number of alleles of one of them. We refer to this as allele 1.

In forensic genetics it is recommended to report the weight of the evidence as the likelihood ratio of two competing hypotheses. The standard forensic genetic analysis deals with linking an identified suspect to biological material found at a scene of crime or on the victim. Then the likelihood ratio evaluates the probability of the evidence under the two scenarios that the suspect *did* and *did not* contribute biological material to the stain of interest. Hence, the two hypothesis are exclusive *and* exhaustive. In the case of ancestry analysis, the two hypothesis is also exclusive – the postulated populations of origin will be assumed independent due to their genealogical history. However, they may very well be non-exhaustive, meaning that the likelihood ratio may be misleading. For example, that likelihood ratios of one population compared to other populations in a database are large does not imply that any of the populations are relevant.

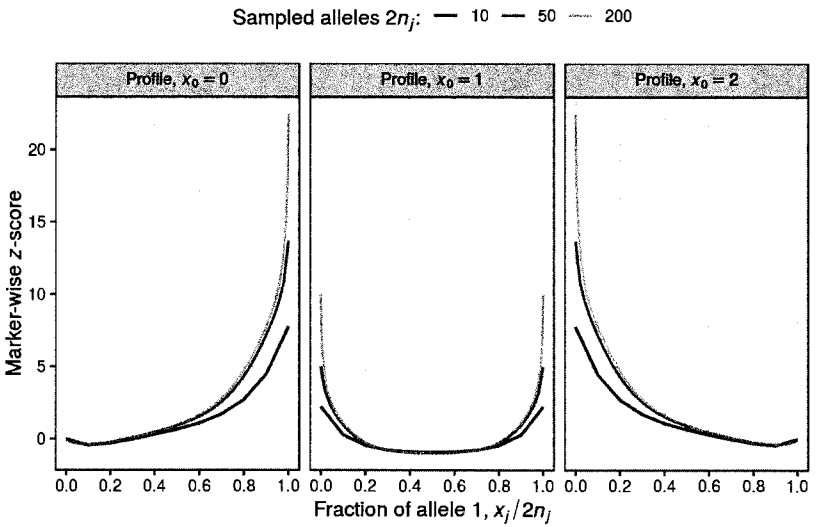
As we are limited to evaluating the probability of the evidence to populations from where we have a population sample, the profile at hand may originate from a population non sampled in the database of reference populations. Hence, by comparing the relative

frequency of the profile, we are unable to assess that none of the populations are sufficiently close to the true population of origin.

To handle this phenomena, we derive a likelihood ratio test (LRT). The LRT is a measure of absolute concordance between a profile and a population, rather than a relative measure of the profile's likelihood in two populations. The LRT is similar to a Fisher's exact test and can be thought of as an outlier test. That is, if a profile is not sufficiently similar to the individuals sampled in the database of reference populations it is declared as an outlier.

Likelihood ratio test based outlier test

Specifically, under the null hypothesis we assume the profile of interest, x_0, and population sample to originate from the same population j. Hence, as each individual has two alleles (one inherited from each parent), the count of allele 1 can be assumed to be binomial with index *2* and $2n_j$ with the same probability parameter, p_j. From the same considerations as Fisher's exact test, we condition on the sufficient statistic under the null hypothesis. The conditional distribution of x_0 given $x_+ = x_0 + x_j$ is hypergeometrical, meaning that we can evaluate the standardized *z*-score to assess whether x_0 is extreme compared to x_j. In the figure below we see how the *z*-score changes with the fraction of allele 1 depending on the state of x_0.



Consequently, we can screen the profile in each of the populations in our database of reference populations. If it rejected in all populations, we refrain from reporting any likelihood ratios as this corresponds to a complete rejection of belonging to any of the sampled populations.

The current AIMs are assumed mutually independent within the reference populations as the markers are located sufficiently far from each other. In addition, statistical tests have not revealed any dependencies violating the assumption of independence. By aggregating over markers, the central limit theorem suggests the resulting quantity is approximately standard normally distributed.

If only a few markers are genotyped or if the majority of the markers are fixated in a given population, the normal approximation may fail. We overcome this using importance sampling and show how exponential tilting gives an efficient proposal distribution. Since x_0 only takes values in $\{0,1,2\}$ the cumulant generating function is easily computed making exponential tilting fast and efficient for importance sampling.

Admixed individuals

The framework outlined above assumed that the parents of the profile of interest came from the same population. However, admixed individuals are frequently occurring, meaning that the parent originate from different populations. By using the EM-algorithm we are able to deal with so-called first-order admixtures, where the parents are non-admixed but from different populations. Future research will focus on generalizing this to the case of second-order admixtures, where the parents themselves may be admixed.

References

Torben Tvedebrink, PS Eriksen, HS Mogensen, N Morling (2018). *Weight of the evidence of genetic investigations of ancestry informative markers*. Theoretical Population Biology 120: 1–10.

Bayesianske hypotesetest[1]

Tom Engsted
Institut for Økonomi, Aarhus Universitet
Fuglesangs alle 4, 8210 Aarhus V.
Email: tengsted@econ.au.dk

December 2018

Abstract: I denne artikel beskrives bayesianske hypotesetests, og der sammenlignes med de klassiske tests. De klassiske tests giver sandsynligheden for data, givet modellen (nulhypotesen), mens de bayesianske tests giver sandsynligheden for modellen, givet data. De to tilgange giver ofte vidt forskellige resultater. Fordele og ulemper ved de to tilgange diskuteres, og der præsenteres empiriske eksempler, der tydeligt illustrerer forskellen på metoderne.

Keywords: Hypotesetest, p-værdi, signifikansniveau, Bayes formel, Bayesfaktor, likelihood ratio test, styrke, andelsparameter, politisk meningsmåling.

1. Indledning.

I Danmark har samfundsforskere traditionelt anvendt klassiske frekvensbaserede statistiske metoder i analyser af empiriske data, og det er også sådanne metoder, der hovedsaligt undervises i i statistik- og økonometriundervisningen på universiteterne. Eksempler på de klassiske metoder er maksimum likelihood estimation af populationsparametre, og hypotesetest på disse parametre baseret på opstilling af nul- og alternativhypotese (H_o og H_1), præspecificeret signifikansniveau (typisk 5%) og beregning af p-værdier.

Brugere af statistiske metoder og læsere af statistiske undersøgelser er ikke altid klar over, hvad metoderne og undersøgelserne reelt viser. Der er eksempelvis mange misforståelser af, hvad p-værdien er for en størrelse. En lav p-værdi er udtryk for, at der i data er stærk evidens imod H_o. Men en

[1] For en uddybende gennemgang og diskussion af metoderne beskrevet i nærværende artikel, se Engsted (2018).

lav p-værdi kan ikke tages som udtryk for, at der er en lav sandsynlighed for at H_o er sand, eller at der er mere evidens for H_1 end for H_o. I den klassiske statistik giver det ikke mening at knytte sandsynligheder til hypoteser, da hypoteser vedrører populationsparametre, der i den klassiske statistik ikke opfattes som stokastiske variable, og p-værdien indeholder ingen information om evidensen for H_1.

Misforståelser og mistolkninger af de klassiske statistiske tests og tilknyttede p-værdier er så udbredte, at det i 2016 fik *The American Statistical Association* til at udsende et officielt dokument om den korrekte brug af sådanne tests: *ASA Statement on Statistical Significance and P-Values* (Wasserstein and Lazar, 2016).

En medvirkende årsag til misforståelserne og mistolkningerne er givetvis, at de klassiske tests på sin vis stiller et bagvendt spørgsmål: Testene bygger på fordelingen af en teststatistik under H_o, og p-værdien giver sandsynligheden for at observere stikprøvens testværdi (eller en mere ekstrem værdi), *givet at H_o er sand.* Alternativhypotesen, H_1, indgår ikke direkte i beregningen af teststatistikken og p-værdien, og et måske mere naturligt spørgsmål at stille er: Givet de observerede data, hvilken af hypoteserne (H_o eller H_1) er da mest sandsynlig?

De klassiske tests besvarer ikke dette spørgsmål (selvom vi ofte lader som om, at de gør). Men der findes en gren af statistikken, hvor dette spørgsmål giver mening og hvor der er udviklet metoder til at besvare spørgsmålet: Bayesiansk statistik. I bayesiansk statistik opfattes populationsparametre som stokastiske variable, og det giver derfor mening at knytte sandsynligheder til hypoteserne H_o og H_1. Bayesianske hypotesetests stiller spørgsmålet: Givet de observerede data, hvad er de relative odds for H_o versus H_1?

I denne artikel beskrives bayesianske hypotesetests, og der sammenlignes med de velkendte frekvensbaserede tests. Klassiske referencer til bayesianske hypotesetests er Jeffreys (1961) og Edwards, Lindman and Savage (1963). Derudover vil jeg inddrage bl.a. Raftery (1986, 1995), Berger and Delampady (1987), Berger and Sellke (1987), Kass and Raftery (1995), Sellke, Bayarri, and Berger (2001), Startz (2014) og Harvey (2017). Simple eksempler vil blive brugt til at illustrere forskellene på klassiske og bayesianske tests, og sluttelig diskuterer jeg fordele og ulemper ved de to tilgange.

2. Bayes formel og Bayes-faktoren.

Fra basal sandsynlighedsregning har vi Bayes formel, der giver den betingede sandsynlighed for hændelsen A, givet hændelsen B, som den omvendt betingede sandsynlighed multipliceret på forholdet mellem de ubetingede sandsynligheder: $P(A \mid B) = P(B \mid A) \cdot \frac{P(A)}{P(B)}$. Antag, at vi ønsker at teste en nulhypotese, H_o, overfor en alternativhypotese, H_1, og at vi på forhånd tildeler de to hypoteser á priori sandsynligheder $P(H_o)$ og $P(H_1) = 1 - P(H_o)$. Lad D være data (stikprøven). Bayes formel indebærer da

$$\begin{aligned} P(H_o \mid D) &= P(D \mid H_o) \cdot \frac{P(H_o)}{P(D)} \\ &= \frac{P(D \mid H_o)P(H_o)}{P(D \mid H_o)P(H_o) + P(D \mid H_1)P(H_1)}, \end{aligned} \tag{1}$$

hvor nævneren i sidste del af udtrykket følger af 'loven om den totale sandsynlighed'.

På tilsvarende vis fås $P(H_1 \mid D) = P(D \mid H_1) \cdot \frac{P(H_1)}{P(D)}$, hvorved den såkaldte *posterior odds ratio* fås som

$$\frac{P(H_o \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_o)}{P(D \mid H_1)} \cdot \frac{P(H_o)}{P(H_1)} = \text{(Bayes-faktor)} \cdot \text{(Prior odds ratio)}. \tag{2}$$

Posterior odds ratio i (2) giver den relative sandsynlighed for H_o versus H_1 som produktet af den relative á priori sandsynlighed (*prior odds ratio* = $\frac{P(H_o)}{P(H_1)}$) og Bayes-faktoren, $BF = \frac{P(D|H_o)}{P(D|H_1)}$, der angiver den relative likelihood. Det ses hermed, hvordan data gennem Bayes-faktoren ændrer vores á priori opfattelse af hypoteserne til en ny á posteriori opfattelse. I sin essens går bayesiansk hypotesetestning ud på at beregne posterior odds ratio som et mål for den relative evidens for de to hypoteser, og dermed som et mål vi kan bruge til at vælge mellem hypoteserne (Jeffreys, 1961), eller som Startz (2014) formulerer det: "*Choosing the more likely hypothesis*". Dette er i modsætning til klassisk hypotesetestning, hvor p-værdien ikke siger noget om evidensen *for* H_o relativt til H_1, men derimod alene noget om evidensen *imod* H_o.

Harvey (2017) benævner $P(H_o \mid D)$ den 'bayesianske p-værdi' ("*Bayesianized p-value*"). Da det naturligvis gælder at $P(H_o \mid D) + P(H_1 \mid D) = 1$,

fås sammenhængen mellem Bayes-faktoren, BF, og den bayesianske p-værdi, fra (2):

$$P(H_o \mid D) = BF \cdot \frac{P(H_o)}{1 + [P(H_o) \cdot (BF - 1)]}. \tag{3}$$

Hovedproblemet i bayesiansk statistik er valget af prior: Hvordan sættes $P(H_o)$? Hvis man ikke har nogen forhåndsformodning eller viden om, at H_o er mere eller mindre sandsynlig end H_1, er standardvalget i bayesiansk analyse: $P(H_o) = P(H_1) = 0.5$, dvs. vi er neutrale overfor de to hypoteser. Hermed forsimples (1) og (3) til:

$$P(H_o \mid D) = \frac{P(D \mid H_o)}{P(D \mid H_o) + P(D \mid H_1)} = \frac{BF}{1 + BF}, \tag{4}$$

og posterior odds ratio i (2) bliver identisk med Bayes-faktoren.

En yderligere komplikation i bayesiansk analyse er beregningen af $P(D \mid H_1)$ når vi har en simpel nulhypotese og sammensat alternativhypotese, hvilket er det mest almindelige. Til illustration af denne komplikation, antag at vi ønsker at teste følgende hypotese om populationsparameteren θ. H_o: $\theta = \theta_o$, H_1: $\theta \neq \theta_o$. I dette tilfælde kræver beregning af $P(D \mid H_1)$ specifikation af en fordeling for θ under H_1. Lad denne fordeling være $f(\theta_1)$. Hermed fås $P(D \mid H_1) = \int P(D \mid \theta_1) f(\theta_1) d\theta_1$. Beregning af dette integrale kan i praksis være udfordrende (diskuteres yderligere nedenfor).

3. Eksempel: Test på en andelsparameter.

Til illustration af den bayesianske testprocedure, antag at vi ønsker at teste, om en given mønt er ægte. Dette kan i den klassiske frekvensbaserede tilgang formuleres enten som et eksakt binomialfordelingstest, eller - hvilket nok er mere almindeligt - som et approksimativt normalfordelingstest på andelsparameteren, θ. Vi kaster mønten $n = 100$ gange og optæller antal 'krone', X. Antag, at vi får $X = 60$. Maksimum likelihood estimatet på andelsparameteren bliver da $\widehat{\theta}_{ml} = \frac{60}{100} = 0.6$. En ægte mønt svarer til H_o: $\theta = 0.5$, mens H_1: $\theta \neq 0.5$ betyder, at mønten er uægte. Baseret på stikprøveestimatet, fås det approksimative normalfordelingstest som

$$Z \text{ statistik} = \frac{\widehat{\theta}_{ml} - \theta_o}{\sqrt{\frac{\theta_o(1-\theta_o)}{n}}} = \frac{0.6 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = 2.0 \sim N(0,1).$$

Den klassiske p-værdi beregnes som $2 \cdot P(Z > 2.0) = 0.046$. Dvs. på det traditionelle 5% signifikansniveau, forkastes H_o, og vi konkluderer at mønten ikke er ægte.

Lad os nu beregne den bayesianske p-værdi og sammenligne med den klassiske. Til det formål kræves der á priori specifikation af $P(H_o)$ og fordelingen af θ under H_1. Hvis vi ikke på forhånd har nogen idé om, at mønten af ægte eller falsk, kan vi sætte $P(H_o) = P(H_1) = 0.5$, hvilket er det neutrale valg. Og da θ naturligt er begrænset til intervallet mellem 0 og 1, kan vi vælge en uniform fordeling over dette interval: $\theta \sim U(0,1)$, dvs. $f(\theta_1) = 1$, jf. Startz (2014). Hermed kan $P(D \mid H_1) = \int P(D \mid \theta_1) f(\theta_1) d\theta_1$ i (4) beregnes som[2]

$$\begin{aligned}
\int P(D \mid \theta_1) f(\theta_1) d\theta_1 &= \int_0^1 \left(\binom{n}{X} \theta_1^X (1-\theta_1)^{n-X} \right) d\theta_1 \\
&= \binom{n}{X} \int_0^1 (\theta_1^X (1-\theta_1)^{n-X}) d\theta_1 \\
&= \frac{n!}{X!(n-X)!} \cdot \frac{X!(n-X)!}{(n+1)!} = \frac{1}{n+1},
\end{aligned}$$

hvor det bemærkes, at $P(D \mid \theta_1)$ er en binomialsandsynlighed, og $\int_0^1 (\theta_1^X (1-\theta_1)^{n-X}) d\theta_1$ er den såkaldte *beta-funktion* med værdien $\frac{X!(n-X)!}{(n+1)!}$ (jf. Berry and Lindgren, 1996, p.355).

Hermed fås fra (4)

$$P(H_o \mid D) = \frac{P(D \mid H_o)}{P(D \mid H_o) + \frac{1}{n+1}}.$$

I eksemplet med $n = 100$ og $X = 60$, fås $P(D \mid H_o) = 0.0108$ (binomialsandsynlighed for 60 'succes' i 100 forsøg med sandsynlighedsparameter = 0.5), hvorved $P(H_o \mid D)$ beregnes til

$$P(H_o \mid D) = \frac{0.0108}{0.0108 + \frac{1}{101}} = 0.522,$$

Posterior odds ratio (= Bayes-faktoren med en neutral prior) bliver lig med

[2]Startz (2014) beregner ikke integralet analytisk, men anvender i stedet numerisk integration. Jeg har tjekket min analytiske beregning med hans numeriske (Figur 4.5 i Startz, 2014), og der er nærmest perfekt overensstemmelse.

$$\frac{P(H_o \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_o)}{P(D \mid H_1)} = \frac{0.0108}{0.0099} = 1.09.$$

Den bayesianske analyse indikerer altså, at der er mere evidens for H_o end for H_1, på trods af, at det klassiske test forkaster H_o på 5% signifikansniveau!

Resultatet kan forekomme paradoksalt. Den klassiske p-værdi på 4.6% indikerer stærk evidens imod H_o, så hvordan kan sandsynligheden for H_o være større end sandsynligheden for H_1? Forklaringen er, at der er endnu stærkere evidens imod H_1. Dette kan også forekomme paradoksalt; for da testet er formuleret som H_o: $\theta = 0.5$ overfor H_1: $\theta \neq 0.5$, dvs. samtlige mulige værdier af θ er indeholdt i de to hypoteser, vil stærk evidens imod H_o så ikke automatisk indebære stærk evidens *for* H_1? Nej, for dét der beregnes, er den *vægtede* likelihood under H_1 (over alle mulige værdier af θ), og denne kan godt være mindre end likelihood under H_o.

De bayesianske sandsynligheder summerer til én: $P(H_o \mid D) + P(H_1 \mid D) = 1$. Dvs. hvis man finder at $P(H_o \mid D)$ er lav, indebærer det nødvendigvis, at $P(H_1 \mid D)$ er høj, og vice versa. Noget tilsvarende gælder ikke i det klassiske test. Den klassiske p-værdi er relateret til $P(D \mid H_o)$. Og selvom man naturligvis også kan beregne $P(D \mid H_1)$ for forskellige θ værdier (hvilket er relateret til et tests styrke, dvs. testets egenskaber under H_1), gælder det ikke at $P(D \mid H_o) + P(D \mid H_1) = 1$, uanset at H_o og H_1 tilsammen indeholder samtlige mulige værdier af θ. $P(D \mid H_o)$ kan være lav samtidig med, at $P(D \mid H_1)$ er endnu lavere. Heraf følger, at en lav p-værdi (eksempelvis mindre end 5%) ikke automatisk kan tages som udtryk for, at data peger på H_1 fremfor H_o, hvad vi ellers normalt lader som om. Da de bayesianske sandsynligheder summerer til én, tilbyder denne metode i den forstand en intuitivt mere naturlig og forståelig procedure til valg mellem H_o og H_1, end den klassiske metode.

4. Simple bayesianske tests.

Som det ses af eksemplet i afsnit 3, kræver Bayes-faktoren og den bayesianske p-værdi evaluering af den vægtede likelihood under H_1, hvilket kræver beregning af et integrale. I simple tilfælde (som i eksemplet) kan dette integrale beregnes analytisk. Men ofte er man i bayesiansk analyse nødsaget til at anvende numerisk integration, og under alle omstændigheder kræves der specifikation af en fordeling for θ under H_1. Disse forhold er med til at begrænse

den praktiske anvendelighed af den bayesianske procedure, sammenlignet med den klassiske testprocedure.

Bayesianske statistikere har derfor udviklet simple metoder til at beregne Bayes-faktoren og den bayesianske p-værdi, $P(H_o \mid D)$, udfra de velkendte klassiske teststatistikker, og hvor der ikke kræves eksplicit specifikation af fordelingen under H_1 eller beregning af et integrale. I det følgende redegøres der for nogle af disse metoder.

4.1 Bayesian Information Criterion (BIC)

Dette kriterium, som også kaldes for Schwarz kriterium, er et ofte anvendt kriterium til valg mellem modeller. Kriteriet har den gode egenskab, at det er konsistent i den forstand, at det asymptotisk vælger den sande model med sandsynlighed én.

Lad LR være den klassiske likelihood ratio test-statistik for modellen under H_o, der er nestet i den mere generelle model under H_1. BIC er da givet som BIC $= (fg \cdot \ln(n)) - LR$, hvor n er stikprøvestørrelsen og fg er antal frihedsgrader. Sammenhængen mellem BIC og Bayes-faktoren er (se Raftery, 1986, 1995; Kass and Raftery, 1995; Startz, 2014):

$$BF_{BIC} = \exp(0.5 \cdot \text{BIC}). \tag{5}$$

I den lineære regressionsmodel med k forklarende variable, og hvor R^2 er determinationskoefficienten ('forklaringsgraden'), er BIC $= (k \cdot \ln(n)) + (n \cdot \ln(1 - R^2))$. I det simple tilfælde, hvor der testes på én koefficient, θ, med det sædvanlige t-test, kan BIC approksimeres ved BIC $= \ln(n) - t^2$, hvor t er t-statistikken for H_o: $\theta = \theta_o$ overfor H_1: $\theta \neq \theta_o$.

Bayesiansk hypotesetestning med anvendelse af BF_{BIC} er blevet ret populær indenfor især sociologisk forskning. Men det er velkendt, at BIC er konservativ i den forstand, at kriteriet har tendens til at vælge modeller med få parametre. I relation til Bayes-faktoren indebærer det, at BIC kræver stærk evidens i stikprøven imod H_o for at BF_{BIC} bliver lav, og BIC har derfor tendes til at favorisere H_o. Som et alternativ til BIC, har statistikere udviklet nedre grænser for Bayes-faktoren, dvs. mål, der i modsætning til BIC har indbygget en tendens til at favorisere H_1. Sådanne nedre grænser gennemgås i afsnit 4.2.

4.2 Nedre grænser for Bayes-faktoren

Edwards, Lindman, and Savage (1963), Berger and Sellke (1987) og Sellke, Bayarri and Berger (2001) udleder nedre grænser for Bayes-faktoren for

forskellige klasser af fordelinger for θ under H_1. Indenfor hver klasse, udgør disse nedre grænser den maksimale evidens imod H_o.

Den nedre grænse for Bayes-faktoren for *alle* tænkelige fordelinger for θ under H_1 er approksimativt

$$BF_A = \exp(-\frac{1}{2}t^2), \tag{6}$$

hvor t som ovenfor er t-statistikken for H_o: $\theta = \theta_o$ overfor H_1: $\theta \neq \theta_o$. Der er dog i litteraturen almindelig enighed om, at denne grænse er for ekstrem. Ikke desto mindre indebærer denne grænse bayesianske p-værdier, der generelt er *højere* end klassiske p-værdier baseret på t-statistikken. Eksempelvis en t-statistik på 1.96, der som bekendt har en klassisk p-værdi på 5%; her fås $BF_A = 0.146$ med anvendelse af (6), og med en neutral á priori sandsynlighed for H_o på $P(H_o) = 0.5$, bliver den bayesianske p-værdi lig med 12.8% ved indsættelse i (4).[3]

Indenfor klassen af *normalfordelinger* for θ under H_1, centreret omkring θ_o, fås den nedre grænse for Bayes-faktoren som

$$BF_N = \sqrt{e} \cdot t \cdot \exp(-\frac{1}{2}t^2), \tag{7}$$

for $t > 1$, og hvor $e \approx 2.718$ er det naturlige grundtal. Med en t-statistik på 1.96 (klassisk p-værdi = 0.05), fås $BF_N = 0.473$ og bayesiansk p-værdi = 0.321 ved indsættelse i (4).

Indenfor klassen af *symmetriske fordelinger* fås den nedre grænse som

$$BF_S = 2\exp(-\frac{1}{2}t^2), \tag{8}$$

for $t > 1$. For $t = 1.96$, fås $BF_S = 0.293$ og bayesiansk p-værdi = 0.227.

Hvis vi yderligere indsnævrer klassen af fordelinger, kan vi betragte klassen af *unimodale symmetriske fordelinger*: Indenfor denne klasse kan den nedre grænse for Bayes-faktoren approksimeres som

$$BF_{US} = -e \cdot p \cdot \ln(p), \tag{9}$$

[3] $BF_A = \exp(-\frac{1}{2}t^2)$ fremkommer på følgende måde: Den minimale værdi af Bayes-faktoren fås når $P(D \mid H_1)$ maksimeres, hvilket sker i $\widehat{\theta}_{ml}$, dvs. når fordelingen for θ under H_1 er koncentreret i maksimum likelihood estimatet. Og da der gælder, at $-2\ln\left(\frac{P(D|\theta_o)}{P(D|\widehat{\theta}_{ml}}\right) = -2\ln(BF_A) \approx t^2$, følger det at $BF_A \approx \exp(-\frac{1}{2}t^2)$. Bemærk ligheden med det klassiske likelihood ratio test (diskuteres nærmere i afsnit 4.4).

for $p < e^{-1} \approx 0.368$, hvor p er den klassiske p-værdi. For $p = 0.05$ ($t = 1.96$), fås $BF_{US} = 0.407$ og bayesiansk p-værdi $= 0.289$.

Det ses, at for alle disse klasser af fordelinger, er den bayesianske p-værdi markant højere end den klassiske p-værdi, hvilket er et velkendt resultat indenfor bayesiansk hypotesetestning: Bayesianske p-værdier er generelt (meget) højere end klassiske p-værdier. Dette er bemærkelsesværdigt, da de bayesianske p-værdier er beregnet udfra Bayes-faktorernes nedre grænser, dvs. med automatisk indbygget stærk evidens imod H_o.

4.3 Hvordan vælges mellem H_o og H_1?

Hvilke værdier af Bayes-faktoren og/eller den bayesianske p-værdi bør generelt lede til forkastelse af H_o? Det kan der ikke gives noget universelt svar på, lige så lidt som der i de klassiske tests findes et 'optimalt' signifikansniveau. Indenfor den klassiske tilgang har man traditionelt valgt et lavt signifikansniveau (typisk 5%) ud fra den betragtning, at der i data skal være meget stærk evidens imod H_o, før man er villig til at forkaste hypotesen, og det kræver et lavt signifikansniveau, svarende til en lav sandsynlighed for at begå en fejl, hvis man forkaster (Type-1 fejl). Man siger også, at H_o er '*the maintained hypothesis*'.

Hvis vi anvender samme tankegang indenfor den bayesianske tilgang, er der umiddelbart to muligheder: Vi kan enten sætte neutrale á priori sandsynligheder for de to hypoteser, dvs. $P(H_o) = P(H_1) = 0.5$, og da vælge at forkaste H_o hvis den bayesianske p-værdi, $P(H_o \mid D)$, er meget lav, eksempelvis 0.05, svarende til en Bayes-faktor på 0.05. Den neutrale prior, $P(H_o) = 0.5$, kan siges at afspejle det klassiske ideal om ikke at være forudindtaget overfor den hypotese man tester og blot lade 'data tale' (dette ideal om 'objektivitet' diskuteres nærmere i afsnit 6). Raftery (1995) og Kass and Raftery (1995) definerer en Bayes-faktor på 0.05 og derunder som værende "stærk" evidens imod H_o netop med henvisning til den klassiske intention bag det traditionelle 5% signifikansniveau.[4]

En anden mulighed er at sætte $P(H_o)$ højt, eksempelvis 0.95, og da forkaste H_o hvis $P(H_o \mid D) < 0.5$. $P(H_o) = 0.95$ afspejler ideen om, at H_o er 'the maintained hypothesis', hvor der skal være meget stærk evidens imod H_o før vi forkaster. Hvis data så fører til at sandsynligheden for H_o

[4]Raftery (1995) og Kass and Raftery (1995) definerer Bayes-faktoren som $BF = \frac{P(D|H_1)}{P(D|H_o)}$ og kalder $BF > 20$ for "stærk" evidens imod H_o. Da min definition af Bayes-faktoren er den reciprokke af deres, svarer dette til en Bayes-faktor på 0.05 og derunder i min notation.

bliver mindre end sandsynligheden for H_1, må vi forkaste H_o til fordel for H_1. I det gennemgående eksempel ovenfor med $t = 1.96$, hvis vi sætter $P(H_o) = 0.95$ og indsætter Bayes-faktorerne (6) - (9) i (3), fås bayesianske p-værdier mellem 0.74 og 0.90, dvs. pænt højere end 0.5. Det ses, at en t-statistik på 1.96, der normalt fører til forkastelse af H_o til fordel for H_1, i en bayesiansk analyse snarere peger på, at H_o er mere sandsynlig end H_1! Den klassiske p-værdi skal generelt være væsentligt lavere end 5% før en bayesiansk analyse peger på H_1 fremfor H_o.

4.4 Sammenligning med likelihood ratio testet

Bayes-faktoren, $BF = \frac{P(D|H_o)}{P(D|H_1)}$ angiver den relative likelihood, en størrelse, der også er velkendt i klassisk statistik. Det klassiske likelihood ratio test fremkommer som $LR = -2\ln\left(\frac{P(D|\theta_o)}{P(D|\widehat{\theta}_{ml})}\right)$. Forskellen mellem BF og LR er, at hvor LR under H_1 evaluerer likelihood-funktionen i maksimum likelihood estimatet, beregnes der i BF et vægtet gennemsnit af likelihood over alle parameterværdier under H_1. Eneste undtagelse er BF_A i (6), hvor $P(D \mid H_1)$ er lig med $P(D \mid \widehat{\theta}_{ml})$, jf. fodnote 3. En yderligere vigtig forskel er, at p-værdien for det klassiske LR test alene beregnes udfra teststørrelsens fordeling under H_o, hvilket ikke er tilfældet for BF i bayesiansk hypotesetestning. Endelig tillader Bayes-faktorer sammenligning af ikke-nestede modeller, hvor modellerne i det klassiske LR test skal være nestede.

4.5 Styrke

Som det fremgår af (1) - (4), indgår sandsynligheden for data under H_1, $P(D \mid H_1)$, direkte i bayesianske tests: Den bayesianske p-værdi afhænger direkte af både $P(D \mid H_o)$ og $P(D \mid H_1)$. I klassiske tests indgår testets egenskaber under H_1 gennem beregning af testets styrke, men i de fleste empiriske undersøgelser indenfor samfundsvidenskab spiller styrken af de anvendte tests en underordnet rolle sammenlignet med p-værdien, når der konkluderes på baggrund af testene. Det er ikke sjældent at se empiriske videnskabelige artikler og undersøgelser, hvor de anvendte tests styrkeegenskaber overhovedet ikke nævnes. I bayesianske tests, derimod, optræder de to hypoteser, H_o og H_1, på lige fod.

I forlængelse af dette, kan vi se lidt nærmere på sammenhængen mellem et klassisk tests styrke og sandsynligheden for at H_o er sand, hvis testet forkaster H_o. Intuitivt vil man måske umiddelbart tænke, at man har et stærkt og troværdigt resultat, hvis man er i stand til at forkaste H_o med et test, der har lav styrke. Men i virkeligheden forholder det sig modsat.

Jo lavere styrke, desto *større* er sandsynligheden for at man begår en fejl, hvis man forkaster med det klassiske test. Dette følger af Bayes formel, der indebærer, at med en neutral prior, $P(H_o) = P(H_1) = 0.5$, gælder (jf. Startz, 2014):

$$P(H_o \mid H_o \text{ forkastes}) = \frac{\alpha}{\alpha + (1 - \beta)},$$

hvor α er testets størrelse (size), dvs. signifikansniveauet (den maksimale sandsynlighed for en Type-1 fejl), og β er sandsynligheden for en Type-2 fejl, dvs. $1 - \beta$ er testets styrke. Det ses, at jo lavere styrke, desto større er sandsynligheden for en '*false discovery*', hvis man forkaster H_o.

5. Eksempel: Politisk meningsmåling.

Diverse analyseinstitutter udsender jævnligt gennem medierne politiske meningsmålinger, der kommenteres flittigt af politikere, journalister og politiske kommentatorer. Målingerne er typisk baseret på stikprøver med omkring 1000 respondenter, og i de fleste af målingerne rapporteres skøn over den statistiske usikkerhed. Eksempelvis Greens Analyseinstituts målinger, der hver måned publiceres i dagbladet Børsen. Disse målinger viser en kolonne med den "statistiske usikkerhed", der for hvert politiske parti er beregnet som

$$1.96\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta}}{n}},$$

hvor $\widehat{\theta}$ er andelen, der i målingen siger de vil stemme på det pågældende parti. Det ses, at den statistiske usikkerhed måles som i et klassisk 95% konfidensinterval. Når kommentatorer kommenterer på, hvorvidt et parti er gået frem eller tilbage i forhold til sidste måling eller sidste folketingsvalg, sker det ofte med eksplicit inddragelse af usikkerheden beregnet på denne måde, dvs. i realiteten baseret på, hvorvidt ændringen er statistisk signifikant i klassisk statistisk forstand og på et 5% signifikansniveau.

Et konkret eksempel er målingen i Børsen den 2. marts, 2015. I denne måling, der er baseret på $n = 1173$ respondenter, opnår Enhedslisten en stemmeandel på 8.4%, og den statistiske usikkerhed oplyses at være 1.6% ($= 1.96\sqrt{0.084(1 - 0.084)/1173}$). Ift. til folketingsvalget den 15. september 2011, hvor partiet fik 6.7% af stemmerne, er partiet altså gået 'signifikant'

frem. Et klassisk test (á la testet i møntkasteksemplet i afsnit 3) på H_o: $\theta = 0.067$ overfor H_1: $\theta \neq 0.067$, giver en Z statistik på 2.33 med en p-værdi på 0.02.

Tabel 1 viser bayesianske tests på samme hypotese, med anvendelse af Bayes-faktorerne fra afsnit 3 og 4 (BF_U er Bayes-faktoren fra afsnit 3, hvor $\theta \sim U(0,1)$ under H_1), og med to forskellige á priori sandsynligheder for $P(H_o)$: 0.50 og 0.95. Det ses, at med en neutral prior ($P(H_o) = 0.5$), er der ingen af de bayesianske p-værdier, der kommer under 0.05. Den laveste p-værdi er 0.062 med anvendelse af BF_A. Men som forklaret i afsnit 4, er der almindelig enighed om, at denne Bayes-faktor giver en for stærk favorisering af H_1.

		$P(H_o \mid D)$
$BF_U = 3.522$	$P(H_o) = 0.50$	0.779
	$P(H_o) = 0.95$	0.985
$BF_{BIC} = 2.269$	$P(H_o) = 0.50$	0.694
	$P(H_o) = 0.95$	0.977
$BF_N = 0.254$	$P(H_o) = 0.50$	0.202
	$P(H_o) = 0.95$	0.828
$BF_{US} = 0.213$	$P(H_o) = 0.50$	0.175
	$P(H_o) = 0.95$	0.802
$BF_S = 0.132$	$P(H_o) = 0.50$	0.117
	$P(H_o) = 0.95$	0.715
$BF_A = 0.066$	$P(H_o) = 0.50$	0.062
	$P(H_o) = 0.95$	0.556

Tabel 1: Bayesianske tests på H_o: $\theta = 0.067$ (valgresultatet i 2011), overfor H_1: $\theta \neq 0.067$, for Enhedslisten i Greens meningsmåling, marts 2015. (Kilde: Børsen, 2. marts, 2015).

Sætter vi i stedet $P(H_o) = 0.95$ som en afspejling af, at H_o er vores 'maintained hypothesis', viser Tabel 1, at alle bayesianske p-værdier er større

end 0.50. Meningsmålingen giver altså ikke tilstrækkelig stærkt grundlag for at ændre vores á priori opfattelse, nemlig at tilslutningen til Enhedslisten ikke har ændret sig ift. folketingsvalget i 2011.

For dette parti giver den bayesianske analyse hermed en anden konklusion end den klassiske analyse. For andre partier i meningsmålingen er der dog overensstemmelse mellem de to metoder. Eksempelvis for partiet Venstre, der i målingen opnår 21.4%, en tilbagegang fra 26.7% ved valget i 2011. En nulhypotese om ingen ændring forkastes klart med både det klassiske test og de bayesianske tests (uanset valg af Bayes-faktor). Samtlige p-værdier er mindre end 0.01.

6. Fordele og ulemper ... og lidt subjektiv causeren.

De fleste samfundsforskere er opflasket med de frekvensbaserede statistiske metoder, hvor klassiske hypotesetests spiller en essentiel rolle. Vi beregner p-værdier og tester rutinemæssigt hypoteser på 5% signifikansniveau. Fordelen ved dette er, at vi alle har et fælles metodemæssigt udgangspunkt med nogen fælles standarder, hvilket gør det lettere at forstå og fortolke egne og andres empiriske undersøgelser. Som empirikere opfatter vi de klassiske metoder som et nyttigt praktisk redskab til at uddrage information fra data, men med mindre vi er fagstatistikere, eller på anden måde har tilegnet os substantiel statistisk indsigt, tænker vi som oftest ikke dybere over de metodemæssige, videnskabsteoretiske og filosofiske aspekter af metoderne, og vi kan være blinde for de problemer og begrænsninger, der kendetegner metoderne.

Den klassiske p-værdi fortæller noget om evidensen i data *imod* H_o, men siger i sig selv intet om evidensen *for* H_o relativt til H_1. Og en lav p-værdi udelukker ikke, at der er stærkere evidens imod H_1 end imod H_o. Bayesianske statistikere har i mange år fremført, at det er den måde der betinges på i den klassiske tilgang, der er problemet. H_o testes grundlæggende ved at spørge: Hvad er sandsynligheden for data, givet H_o? Men et mere naturligt spørgsmål at stille er: Hvad er sandsynligheden for H_o, givet data? Vi har to hypoteser, H_o og H_1, og ønsker den relative evidens i data for de to hypoteser. Bayes-faktoren og den bayesianske p-værdi besvarer dette spørgsmål, da der i de bayesianske tests betinges på data og ikke på H_o. Men klassiske statistikere har traditionelt afvist denne tilgang, da hypoteser vedrører populationsparametre og disse i klassisk statistik ikke opfattes som stokastiske

variable, hvorved det ikke giver mening at tale om sandsynligheden for H_o eller de relative odds for H_o versus H_1.

Klassiske statistikere har også traditionelt opponeret mod den bayesianske tilgangs afhængighed af priors. For at beregne $P(H_o \mid D)$ vha. Bayes formel, kræves der specifikation af en á priori sandsynlighed, $P(H_o)$. Hvordan fastsættes denne sandsynlighed objektivt? Bayesianere ser ikke noget problem her, for i bayesiansk statistik opfattes sandsynligheder i forvejen som subjektive. For bayesianere udtrykker sandsynlighed graden af tro ('*degree of belief*') snarere end grænseværdien af en relativ frekvens, som i den klassiske frekvensbaserede statistik, og denne grad af tro er i sagens natur forskellig fra person til person. Men uanset dette, kan man jo altid sætte $P(H_o) = P(H_1) = 0.5$, hvilket er det neutrale ('objektive') valg, hvis man ikke er villig til - eller i stand til - at udtrykke en personlig á priori holdning til H_o.

Det har været den almindelige opfattelse, at de klassiske statistiske metoder er mere objektive end de bayesianske metoder netop fordi de klassiske metoder ikke direkte afhænger af priors. De klassiske metoder lader i højere grad 'data tale for sig selv'. Bayesianere kalder denne opfattelse en illusion. For det første træffer vi mange valg med subjektive elementer inden vi når til den egentlige hypotesetestning: Valg af datakilder, analyseperiode, forklarende variable, funktionsform, estimationsmetode, osv. For det andet er valget af signifikansniveau i det klassiske test subjektivt, da der ikke findes et optimalt signifikansniveau. Det konventionelle valg på 5% er arbitrært og stammer i virkeligheden fra en gammel tommerfingerregel, som Fisher (1925) formulerede, og som han aldrig selv opfattede som en fast dogmatisk regel. Efterfølgende har man typisk begrundet valget af et lavt signifikansniveau med, at H_o udtrykker 'the maintained hypothesis', hvor der kræves stærk evidens imod hypotesen før man er villig til at forkaste den. Men deri ligger der i sig selv en subjektiv forhåndsvurdering af H_o. Startz (2014) argumenterer for, at klassiske tests i valget af signifikansniveau har indbygget en (skjult) prior, og han viser eksempler, hvor han udleder den implicitte $P(H_o)$ fra de klassiske tests. Disse implicitte priors er ofte langt fra neutrale! Alternativt kan valget af et lavt signifikansniveau i den klassiske analyse være udtryk for, at man vægter Type-1 fejl højere end Type-2 fejl. Men dette kræver opstilling af en tabsfunktion for de to fejltyper, der direkte relateres til den foreliggende problemstilling. Noget sådant sker meget sjældent i klassiske empiriske analyser.

Og vedrørende argumentet om at H_o udtrykker 'the maintained hypoth-

esis' og derfor kræver et lavt signifikansniveau, gør vi i virkeligheden ofte det modsatte når vi formulerer og tester vores modeller. Ikke sjældent er vores 'maintained hypothesis' i virkeligheden H_1. Økonomer udvikler teoretiske modeller, der indebærer, at X påvirker Y. Dét er vores 'maintained hypothesis' vi ønsker at teste. Men i formuleringen af testet lader vi typisk H_o være, at X *ikke* påvirker Y (eksempelvis H_o: $b = 0$ overfor H_1: $b \neq 0$ i regressionen $Y = a + bX + u$), og alligevel vælger vi oftest at teste på det traditionelle 5% signifikansniveau.

Bayesiansk statistik foregiver ikke at være objektiv, men forsøger at være så åben og eksplicit om de valgte subjektive priors som muligt. Dette udelukker naturligvis ikke, at der kan ske mistolkninger og misbrug af bayesiansk statistik. Flere af de løftede pegefingre, som Wasserstein and Lazar (2016) på vegne af *The American Statistical Association* retter mod anvendelser af de klassiske statistiske metoder, kan givetvis også rettes mod mange bayesianske anvendelser.

En konkret komplikation i bayesiansk analyse, der ikke findes i den klassiske analyse, er specifikation af en fordeling for modellens parametre under H_1, og tilhørende beregning af den vægtede likelihood under H_1. Dette kræver ofte anvendelse af numerisk integration. Og selvom moderne computere og tilgængeligt software har gjort dette problem mindre, er det givetvis en medvirkende årsag til, at bayesianske metoder fortsat er relativt ubenyttede i empirisk samfundsforskning. Men de simple bayesianske tests gennemgået i afsnit 4 kræver ikke en eksplicit specifikation af parametrenes fordeling under H_1. Disse tests kan beregnes ret simpelt udfra de velkendte klassiske teststatistikker, og udgør i den forstand et kompromis mellem de to tilgange.

Der er fordele og ulemper ved både klassisk og bayesiansk statistik, og - som vi har set - fører de to tilgange ofte til ret forskellige konklusioner. Empiriske samfundsforskere har de senere år i stigende omfang fået øjnene op for de muligheder den bayesianske tilgang tilbyder. Eksempelvis anbefaler Startz (2014) og Harvey (2017), at praktiske økonometrikere og empiriske finansieringsøkonomer fremover rapporterer bayesianske tests på lige fod med de sædvanlige klassiske tests når der fremlægges videnskabeligt arbejde. Tiden må vise, om den bayesianske tilgang efterhånden får samme plads i værktøjskassen som den klassiske frekvensbaserede tilgang.

Referencer.

Berger, J.O. and M. Delampady (1987): Testing precise hypotheses. *Sta-*

tistical Science 2, 317-352.

Berger, J.O. and T. Sellke (1987): Testing a point null hypothesis: The irreconcilability of P values and evidence. *Journal of the American Statistical Association* 82, 112-122.

Berry, D.A. and B.W. Lindgren (1996): *Statistics: Theory and Methods* (2nd edition). Duxbury Press.

Edwards, W., H. Lindman and L.J. Savage (1963): Bayesian statistical inference for psychological research. *Psychological Review* 70, 193-242.

Engsted, T. (2018): Frekvensbaserede versus bayesianske metoder i empirisk økonomi. *Economics Working Papers 18-07*, Aarhus Universitet

Fisher, R.A. (1925): *Statistical Methods for Reseach Workers.* Oliver and Boyd Ltd., Edinburgh.

Harvey, C.R. (2017): Presidential address: The scientific outlook in financial economics. *Journal of Finance* 72, 1399-1440.

Jeffreys, H. (1961): *Theory of Probability* (3rd ed.). Oxford University Press.

Kass, R.E. and A.E. Raftery (1995): Bayes factors. *Journal of the American Statistical Association* 90, 773-795.

Raftery, A.E. (1986): Choosing models for cross-classifications. *American Sociological Review* 51, 145-146.

Raftery, A.E. (1995): Bayesian model selection in social research. *Sociological Methodology* 25, 111-163.

Sellke, T., M.J. Bayarri, and J.O. Berger (2001): Calibration of p values for testing precise null hypotheses. *The American Statistician* 55, 62-71.

Startz, R. (2014): Choosing the more likely hypothesis. *Foundations and Trends in Econometrics* 7, 119-189.

Wasserstein, L.R. and N.A. Lazar (2016): The ASA's statement on p-values: Context, process, and purpose. *American Statistician* 70, 129-133.

A weighted median approach to correct for dropout in randomised clinical trials

Mia Klinten Grand[1,2] and Volkert Dirk Siersma[1]

[1]The Research Unit for General Practice and Section of General Practice, University of Copenhagen, Denmark

[2]Section of Biostatistics, University of Copenhagen, Denmark

In randomised clinical trials it is often the case that the outcome is not observed for all subjects at the end of the study due to dropout or death. One way to account for dropout is to correct the mean estimate by using inverse probability of censoring weights. When the outcome is continuous but heavily skewed the mean estimate becomes very sensitive to outliers. In some instances an analysis of the mean may be alleviated by a transformation of the outcome. However, most of the standard transformations only work well for positively right skewed distributions. An alternative is to perform an analysis on the median, which is more robust to outliers and applicable to a wider range of distributions. We propose a weighted median approach to estimate the median of the outcome among those still alive at the end of the study. The performance of the method is investigated in a simulation study and applied to a randomised clinical trial on Type 2 Diabetes patients.

Keywords: quantile regression; randomized clinical trial; inverse probability weighting; differential dropout; missing

1 Introduction

In a regular randomised clinical trial the objective is to compare a new treatment to the standard treatment. If the outcome of the trial is a continuous biomarker the target estimand is most often the average treatment effect. Without missing observations the average treatment effect can be estimated by taking the difference between the mean outcome in the two groups. However, it is often the case that the outcome is not observed for all subjects at the end of the study due to dropout or death. It is well known that ignoring missing observations in general can lead to biased

estimates, if the observations are not missing completely at random [Little and Rubin, 1987]. In a randomised clinical trial missing observations become a problem when they are not independent of the randomisation.

If outcomes are missing at random only due to dropout, there are a number of different approaches to correct for the bias caused by dropout. One such approach is inverse probability weights, where the observed outcomes are weighted so that they also represent those that dropped out of the study. In this setting the target estimand would still be to estimate the average treatment effect.

If outcome are missing due to both dropout and death we have to be more careful [Dufouil et al., 2004, Kurland and Heagerty, 2005]. Dufouil et al. [2004] argued that outcomes that are missing due to death should be handled separately from dropout. In this setting they advocate to let the target estimand be the average treatment effect given that the patient is alive at the end of the study. Their argument is that since we cannot eradicate death this estimand is relevant for many interest groups. It is certainly relevant from a health care manager perspective. Another argument is that if we insist on estimating the average treatment effect, then we would have to extrapolate the outcome beyond death by letting the living represent the dead. This will often be an ambitious endeavour, unless death does not depend on the biomarker outcome or if the continuous outcome is something which is also well defined after death [Kurland et al., 2009]. If attainable, this estimand is typically of interest for researchers interested in aetiology.

If the outcome distribution is heavily skewed or has heavy tails the mean estimate becomes very sensitive to outliers. Sometimes this can be alleviated by transforming the outcome, but it is not always possible to find a suitable transformation. An alternative is to consider the median, since it is more robust against outliers and applicable to a wider range of distributions. Median regression, or more generally quantile regression, was proposed by Koenker and Bassett Jr [1978]. A number of different approaches have been developed to handle missing data in quantile regression. In particular inverse probability weights have previously been used to account for dropout in quantile regression models for longitudinal data using a parametric model for the probability weights [Lipsitz et al., 1997, Yi and He, 2009]. Chen et al. [2015] used inverse probability weights to account for both missing covariates and outcome using a nonparametric model for the probability weights.

If the outcomes are not missing due to death, the target estimand is the median treatment effect. Without dropout, it is simply estimated by taking the difference between the median in the two groups.

In this paper we describe how to estimate the quantile treatment effect given that the patient is alive at the end of the study, when the outcome distribution has heavy tails and are subject to

dropout and death, using inverse probability weights. In Section 2 we describe the method and explain how it differs from the setting without missing observations and when observations are only missing due to dropout. In Section 3 we investigate the finite sample performance of the method under different scenarios and compare it to an naive analysis using only complete cases. We applied the method to a randomised clinical trial on Type 2 Diabetes patients in Section 4, and we end with a summary and discussion in Section 5.

2 Method

Consider a setting where we have a standard randomized clinical trial with n subjects. Let $A_i \in \{0, 1\}$ denote the treatment that was assigned to subject i and let $\boldsymbol{Y}_i$ denote the continuous outcome observed at the end of the trial. Let $\boldsymbol{Z}_i$ denote a vector of covariates measured at baseline. We will refer to treatment 0 as the standard treatment and to 1 as the new intervention treatment. We assume that there is a linear relation between the treatment and the τth quantile of the outcome

$$Q_{\boldsymbol{Y}_i}(\tau|A_i) = \alpha_0(\tau) + \alpha_1(\tau)A_i \ ,$$

Hence, $\alpha_0(\tau)$ is the τth quantile with the standard treatment and $\alpha_1(\tau)$ is the quantile treatment effect. When $\tau = 0.5$, then $\alpha_1(0.5)$ is of course the median treatment effect. The parameters $\boldsymbol{\theta}(\tau) = [\alpha_0(\tau), \alpha_1(\tau)]^\top$ can be estimated by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} l_\tau\big(\boldsymbol{Y}_i - \alpha_0(\tau) - \alpha_1(\tau)A_i\big) \ ,$$

where $l_\tau(x) = x\big(\tau - I(x < 0)\big)$ for $\tau \in (0, 1)$ [Koenker and Bassett Jr, 1978]. The estimator $\hat{\boldsymbol{\theta}}(\tau)$ is also the solution to the estimating equations

$$\frac{1}{n}\sum_{i=1}^{n} g_\tau\big(\boldsymbol{Y}_i, A_i, \boldsymbol{\theta}(\tau)\big) \approx 0 \ ,$$

where $g_\tau(\boldsymbol{Y}_i, A_i, \boldsymbol{\theta}(\tau)) = [1, A_i]^\top\big(I(\boldsymbol{Y}_i - \alpha_0(\tau) - \alpha_1(\tau)A_i \leq 0) - \tau\big)$. The estimator $\hat{\boldsymbol{\theta}}(\tau)$ is consistent and asymptotically normal.

2.1 Dropout only

Now consider a setting where subjects may drop out before the end of the trial. Let ρ_i denote whether subject i dropped out ($\rho_i = 0$) or was observed ($\rho_i = 1$) at the end of the trial. Let $p_i = P(\rho_i = 1|\boldsymbol{Z}_i, A_i)$ denote the probability of not dropping out. Chen et al. [2015] proposed to use inverse probability weights to account for missing observations using a nonparametric estimator

for the weights. The estimated weights $1/\hat{p}_i$ were then used to weight each subject's contribution to the estimating equations, which in our setting corresponds to

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\rho_i}{\hat{p}_i}g_\tau\big(\boldsymbol{Y}_i, A_i, \boldsymbol{\theta}(\tau)\big) \approx \mathbf{0} \ , \tag{1}$$

Notice that subjects that dropped out have a contribution of 0, but they are still included in the sample size n. Chen et al. [2015] provides conditions under which the estimator $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normal.

Instead of using a nonparametric model for the weights another option is to use a parametric model. We are using logistic regression for the weights in both the simulations and the application.

Nonparametric bootstrap is convenient to use the inference, since it makes it straightforward to account for the additional uncertainty of having to estimate the weights. To calculate the bootstrapped confidence intervals we create m data sets, each of size n, by sampling with replacement from the original data set. For the lth bootstrapped data set, we estimate the weights and plug them into (1) to obtain the estimates $\boldsymbol{\theta}^l(\tau)$. The empirical variance of the estimates from the bootstrapped data sets $\big(\boldsymbol{\theta}^1(\tau), \ldots, \boldsymbol{\theta}^m(\tau)\big)$ is used to estimate the variance of $\hat{\boldsymbol{\theta}}(\tau)$. The confidence intervals are the calculated the usual way.

It is straightforward to implement the method in R. We used the `rq` function from the `quantreg` package that has a weight argument. However, the nonparametric bootstrap has to be implemented by hand.

2.2 Dropout and death

Consider a setting where subjects may drop out or die before the end of the trial. In addition to the drop out indicator ρ_i, let δ_i denote if subject i was alive ($\delta_i = 1$) or had died ($\delta_i = 0$) before the end of the trial. Let $R = \{i|\delta_i = 1\}$ denote the subset of the N subjects that were still alive at the end of the trial.

Analogous to Dufouil et al. [2004] suggestion for mean regression under dropout and death, we would like to estimate the quantile treatment effect among the survivors. To account for dropout we use the probability of being observed among the survivors $p_i^o = P(\rho_i = 1|\delta_i = 1, Z_i, A_i)$. The parameters $\boldsymbol{\theta}(\tau)$ are estimated by solving

$$\frac{1}{N}\sum_{i\in R}\frac{\rho_i}{\hat{p}_i^o}g_\tau\big(\boldsymbol{Y}_i, A_i, \boldsymbol{\theta}(\tau)\big) \approx \mathbf{0} \ . \tag{2}$$

The parameter $\alpha_1(\tau)$ is the quantile treatment effect in the cohort of survivors.Estimation of the weights and confidence intervals can be carried out as described above. The difference is however

that only subjects that are still alive at the end of the study are used. An underlying assumption is that dropout does not influence the risk of death and for some applications this might not be realistic. We investigated how a violation of this assumption affects the estimates in the simulation study. We will refer to this as the *weighted quantile given survival* approach. A naive alternative to this approach is to ignore dropout and attempt to estimate the target parameter by only using the complete cases. We will refer to this as the *complete cases* approach.

If we would insist on estimating the quantile treatment in this setting we would have to accept the premise that the living can be used represent the dead. We could then use the inverse of the probability of being observed, $p_i^{\rho,\delta} = P(\max(\rho_i, \delta_i) = 1 | Z_i, A_i)$, as weights. The parameters $\boldsymbol{\theta}(\tau)$ are estimated by solving

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\rho_i \delta_i}{\hat{p}_i^{\rho,\delta}} g_\tau\big(\boldsymbol{Y}_i, A_i, \boldsymbol{\theta}(\tau)\big) \approx \mathbf{0} \ .$$

The parameter $\alpha_1(\tau)$ is then the quantile treatment effect in an immortal cohort.We will refer to this as *weighted median among immortals* approach.

3 Simulation study

We investigated the finite sample properties of the method in a small simulation study. We considered two scenarios, one where dropout did not increase the risk of death and one where it did. In each scenario we compared the performance of the weighted median given survival to the complete cases approach. We also looked at the performance of the weighted median among immortals approach, although it is not directly comparable to the other two since the target parameter is different.

We simulated data from a setting where the new treatment increased the outcome, and decreased the risk of dropout and death, compared to the standard treatment. We included one continuous baseline covariate, where increasing values decreased the outcome and increased the risk of dropout and death. We simulated data for as single subject i as follows

$$\begin{aligned} A_i &\sim \mathcal{B}(1, 0.5) \\ Z_i &\sim \mathcal{N}(0, 1) \\ Y_i &= A_i - 5Z_i + \varepsilon_i, \text{ where } \varepsilon_i \sim t(3) \end{aligned}$$

Here, $\mathcal{B}$ denotes the binomial distribution, $\mathcal{N}$ the normal distribution and t the Student's t-distribution. Dropout was simulated from a logistic regression model, where A_i (OR= 0.75) was associated with decrease and Z_i (OR= 2) was associated with an increase compared to the reference group

($p = 0.10$). Death was also simulated from a logistic regression model. In the first scenario only A_i (OR= 0.9) and Z_i (OR= 2) had an effect on the risk of death compared to the reference group ($p = 0.25$). In the second scenario dropout (OR= 5) also had an effect on the risk of death. We considered samples sizes of 500 and 1000 subjects and we used 1000 repetitions. The percentage of dropout among the survivors were around 10% and the percentage of deaths were around 25%.

We applied the three approaches to each simulated data set to calculate the quantile treatment effect for $\tau = 0.25, 0.5, 0.75$. In the complete cases approach only the patients who were alive and observed at follow-up were used. In the weighted quantile given survival approach all patients were used to calculate the dropout weights. However, only those alive at follow-up were used to estimate the quantile treatment effect. In the weighted quantile among immortals approach all patients were used. Both to calculate the weights for dropout and death combined and to estimate the quantile treatment effect. Confidence intervals were calculated using a 1000 bootstrap samples.

The performance was evaluated based on the bias, variance, root mean squared error (rmse) of the estimate and coverage percentage of the 95% bootstrapped confidence interval.

3.1 Simulation results

Table 1 shows the performance of the estimated quantile treatment effects in the first scenario where dropout did not change the risk of death. In general the bias of the weighted median given survival was reasonable even with half the sample size, and even smaller actual sample size due to dropout and death. The variance and rmse improved with increasing sample size as expected. The method also seems to perform well for the first and third quartile compared to the median, but we would expect worse performance for smaller sample sizes and τ close to 0 or 1.

Compared to the weighted median given survival the complete cases approach underestimated the true treatment effect in this scenario. If the effect of the covariate Z_1 on the outcome had had opposite sign, then the complete cases approach would have overestimated the true treatment effect. The variance was smaller for the complete cases approach although the weighted median given survival approach utilized more observations. This is probably because it accounts for the uncertainty of the having to estimate the weights. The rmse was overall larger for the complete cases approach, except for the first quartile. In general the coverage probability was close to the nominal level of 0.95.

Table 2 shows the performance of the estimated quantile treatment effects in the second scenario where dropout changed the risk of death. In this scenario the assumptions of the method were violated. Since those that dropped out also tended to die we would expect the estimates from the weighted median given survival and complete cases approach to be more alike, which also seemed to be the case. However, the weighted median given survival approach is no longer able to estimate

the target parameter, which is why the bias was larger compared to Table 1. The complete cases approach has the same problem, but in this scenario it improved in terms of bias, because the bias caused by dropout was less dominating.

4 Application

We used data from the randomized trial Diabetes Care in General Practice [de Fine Olivarius et al., 2001] to illustrate the method. Newly diagnosed Type 2 Diabetes patients were randomized to routine (standard treatment) or structured personal care (new treatment). One of the secondary outcomes were a urinary biomarker of RNA oxidation 8-oxo-7,8-dihydroguanosine adjusted for creatinine, which was measured after around six years of follow-up. Increased levels of RNA oxidation has been associated with higher mortality risk in Type 2 Diabetes patients, both in this trial [Broedbaek et al., 2011] and an independent cohort [Kjær et al., 2017]. The objective of this analysis was to investigate if there is an effect of the new intervention on RNA oxidation compared to the standard treatment. The hope was that if the new treatment lowered RNA oxidation compared to the standard treatment, then it would ultimately improve long term survival.

Out of the 1263 patients, that were included at the beginning of the trial, there were 129 who dropped out without dying and 293 who died before the six year follow-up measurement (Table 3). Although the percentage of dropout is similar in the two treatment groups, it does not automatically imply that we could have ignored dropout. Among the complete cases the RNA oxidation had a right skewed distribution with heavy tails (Figure 1). We first conducted a naive analysis of the data, where we only used the complete cases. When we used a t-test on the log transformed RNA oxidation, the conclusion changed if outliers were removed. When we used a Wilcoxon rank-sum test it rejected that the two distributions were equal. Hence, the situation called for a robust effect measure, where missing observations could be appropriately adjusted for.

We applied all three approaches described in Section 2 to the data. The different weights to correct for missing observations were estimated using logistic regression adjusted for the baseline variables; age, sex, treatment. The 95% confidence intervals were calculated using 1000 bootstrap samples. In addition, it seemed like a reasonable assumption that dropout from the trial did not change the risk of death, or at least only to a very small degree.

4.1 Application results

Table 4 shows the estimated quantile treatment effects and the corresponding 95% confidence interval from applying the three different approaches. The treatment effect is calculated at the median ($\tau = 0.5$), and at the first ($\tau = 0.25$) and third ($\tau = 0.75$) quartile. Although the results are

Table 1: Summary statistics from the three approaches applied to the scenario where dropout did not change the risk of death. It shows the sample size (n), the true quantile effect $\alpha_1(\tau)$ (true), the bias, variance (var) and rmse of the estimator $\hat{\alpha}_1(\tau)$ and the coverage probability (cp).

	$\tau = 0.25$					$\tau = 0.50$					$\tau = 0.75$				
Approach	true	bias	var	rmse	cp	true	bias	var	rmse	cp	true	bias	var	rmse	cp
$n = 500$															
Given Survival	0.92	0.003	0.609	0.780	0.95	0.96	−0.017	0.457	0.676	0.96	0.94	−0.021	0.532	0.729	0.95
Complete cases	0.92	−0.074	0.553	0.747	0.95	0.96	−0.119	0.442	0.675	0.95	0.94	−0.077	0.527	0.729	0.95
Among immortals	0.98	−0.055	0.713	0.846	0.95	1.00	−0.049	0.507	0.713	0.95	0.99	−0.048	0.514	0.718	0.95
$n = 1000$															
Given survival	0.92	0.022	0.276	0.525	0.95	0.96	−0.012	0.236	0.486	0.95	0.94	0.006	0.269	0.519	0.96
Complete cases	0.92	−0.063	0.256	0.509	0.94	0.96	−0.099	0.222	0.481	0.94	0.94	−0.057	0.272	0.525	0.95
Among immortals	0.98	−0.013	0.362	0.601	0.95	1.00	−0.029	0.262	0.512	0.95	0.99	−0.019	0.256	0.506	0.96

Table 2: Summary statistics from the three approaches applied to the scenario where dropout changed the risk of death. It shows the sample size (n), the true quantile effect $\alpha_1(\tau)$ (true), the bias, variance (var) and rmse of the estimator $\hat{\alpha}_1(\tau)$ and the coverage probability (cp).

	$\tau = 0.25$					$\tau = 0.50$					$\tau = 0.75$				
Approach	true	bias	var	rmse	cp	true	bias	var	rmse	cp	true	bias	var	rmse	cp
$n = 500$															
Given Survival	0.92	−0.033	0.572	0.757	0.95	0.96	−0.060	0.459	0.680	0.94	0.94	−0.033	0.536	0.733	0.95
Complete cases	0.92	−0.070	0.553	0.746	0.95	0.96	−0.110	0.442	0.674	0.95	0.94	−0.064	0.527	0.728	0.95
Among immortals	0.98	−0.055	0.713	0.846	0.95	1.00	−0.049	0.507	0.713	0.95	0.99	−0.048	0.514	0.718	0.95
$n = 1000$															
Given survival	0.92	−0.030	0.263	0.513	0.95	0.96	−0.051	0.228	0.480	0.96	0.94	−0.015	0.271	0.521	0.96
Complete cases	0.92	−0.059	0.256	0.509	0.94	0.96	−0.091	0.222	0.480	0.94	0.94	−0.044	0.272	0.523	0.95
Among immortals	0.98	−0.013	0.362	0.601	0.95	1.00	−0.029	0.262	0.512	0.95	0.99	−0.019	0.256	0.506	0.96

Table 3: Distribution of the baseline variables and causes of missing outcomes in the two treatment groups.

	Standard treatment ($n = 614$)	New treatment ($n = 649$)
Baseline variables		
Age (median (interquartile range))	65.3 (56.3; 73.5)	65.5 (55.3; 74.0)
Sex males (n (%))	326 (53%)	340 (52%)
Causes of missing outcome		
Dropout (n (%))	70 (11%)	59 (9%)
Death (n (%))	147 (24%)	146 (23%)

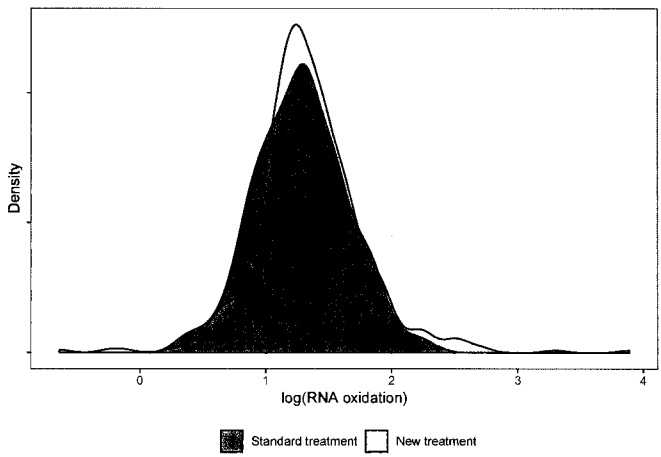


Figure 1: Smoothed density plot of log(RNA oxidation) among the complete cases in each treatment group.

Table 4: Estimated quantile treatment effect ($\alpha_1(\tau)$) and corresponding 95% confidence interval (CI) using different approaches.

Approach	$\alpha_1(0.25)$ (CI)	$\alpha_1(0.50)$ (CI)	$\alpha_1(0.75)$ (CI)
Given survival	0.23 (0.05; 0.41)	0.04 (−0.17; 0.21)	0.08 (−0.28; 0.35)
Complete cases	0.23 (0.04; 0.39)	0.05 (−0.15; 0.22)	0.11 (−0.21; 0.42)
Among immortals	0.21 (0.03; 0.38)	0.02 (−0.23; 0.24)	0.06 (−0.37; 0.39)

somewhat similar for all three approaches the interpretation is inherently different.

The new treatment had a significantly increased first quartile compared to the standard treatment among those that survived. However, there was no difference between the treatments in the median or third quartile. These results are also supported by Figure 1. Although we had hoped that the new treatment would decrease RNA oxidation this seemed not to be the case.

Although dropout was differential the results from the complete cases approach are very similar to the weighted median given survival. One explanation could be that the baseline variables only had a small effect on RNA oxidation after six years. In addition the confidence intervals are in general larger than the weighted median given survival approach.

The estimates from the weighted median among immortals approach suggests that the difference between the two treatments is even smaller in an immortal cohort.

5 Discussion

We proposed to use weighted median regression to correct for dropout, when the outcome is subject to dropout and death. The method can handle a wide range of distributions compare to weighted mean regression and it is robust against outliers.

This work was motivated by data from the randomized trial on newly diagnosed Type 2 Diabetes patients. Our focus have therefore been to estimate the treatment effect in a simple randomized clinical trial, but it would be interesting to extend the method to a more general setting. Often patients are subject to several follow-up visits and it would be a natural next step to extend the method to longitudinal data. If we had known the time of dropout we could extend the model for dropout to an event history model, e.g. a Cox proportional hazards model.

References

K. Broedbaek, V. Siersma, T. Henriksen, A. Weimann, M. Petersen, J. T. Andersen, E. Jimenez-Solem, E. S. Stovgaard, L. J. Hansen, J. E. Henriksen, et al. Urinary markers of nucleic acid

oxidation and long-term mortality of newly diagnosed type 2 diabetic patients. *Diabetes care*, 34(12):2594–2596, 2011.

X. Chen, A. T. Wan, and Y. Zhou. Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association*, 110(510):723–741, 2015.

N. de Fine Olivarius, H. Beck-Nielsen, A. H. Andreasen, M. Hørder, and P. A. Pedersen. Randomised controlled trial of structured personal care of type 2 diabetes mellitus. *BMJ*, 323(7319): 970, 2001.

C. Dufouil, C. Brayne, and D. Clayton. Analysis of longitudinal studies with death and drop-out: a case study. *Statistics in Medicine*, 23(14):2215–2226, 2004.

L. K. Kjær, V. Cejvanovic, T. Henriksen, K. M. Petersen, T. Hansen, O. Pedersen, C. K. Christensen, C. Torp-Pedersen, T. A. Gerds, I. Brandslund, et al. Cardiovascular and all-cause mortality risk associated with urinary excretion of 8-oxoguo, a biomarker for rna oxidation, in patients with type 2 diabetes: A prospective cohort study. *Diabetes care*, 40(12):1771–1778, 2017.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

B. F. Kurland and P. J. Heagerty. Directly parameterized regression conditioning on being alive: analysis of longitudinal data truncated by deaths. *Biostatistics*, 6(2):241–258, 2005.

B. F. Kurland, L. L. Johnson, B. L. Egleston, and P. H. Diehr. Longitudinal data with follow-up truncated by death: match the analysis method to research aims. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(2):211, 2009.

S. R. Lipsitz, G. M. Fitzmaurice, G. Molenberghs, and L. P. Zhao. Quantile regression methods for longitudinal data with drop-outs: application to cd4 cell counts of patients infected with the human immunodeficiency virus. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):463–476, 1997.

R. J. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 1987.

G. Y. Yi and W. He. Median regression models for longitudinal data with dropouts. *Biometrics*, 65 (2):618–625, 2009.

The Pareto distribution. An introduction to theory and some applications

Erik Holst, MSc, Ph.D., HolStat Consulting, erik.holst@teliamail.dk

Introduction

In 1897 the Swiss (Italian born) professor of economics Vilfredo Pareto formulated the so-called Pareto Law $N(x) = Ax^{-\alpha}$, where $N(x)$ is the number of persons in a population having an income $\geq x$. A and α are positive parameters where α is known as Pareto's constant.

The equation $N(x) = Ax^{-\alpha}$ is also known as "the Power law" or "the Power function". Obviously, Pareto's equation $N(x) = Ax^{-\alpha}$ is not a statistical distribution. However, based on the equation a number of statistical distributions have been evolved which have a number of interesting properties and applications.

Since 1970 according to Johnson et al. (1994) the literature on the Pareto distribution has grown perhaps faster than for any other univariate statistical distribution. The production has continued to grow in the years after 1994 and the amount of literature on the Pareto distributions and their applications is now very large. Therefore, it may be expected to find a number of "irregularities" like different names for the same concepts, variable notation etc. in the literature.

The purpose with this paper is to do a little "cleaning" in this mess on the basic level by discussing the original Pareto equation and to treat the Pareto distributions of the first and the second kind.

These two distributions are also denoted as the type 1 and type 2 Pareto distributions. Further we will discuss some of the so-called generalized Pareto distributions. Finally, the discrete Pareto distribution is discussed on the basic level.

Empirical studies of Pareto's law have shown that a number of measurements within very different fields obey the Power law. A large number of examples are given in Wikipedia (https://en.m.wikipedia.org/wiki/Power_law).

The parameters estimated by Pareto resulted in Pareto's law or Pareto's principle which roughly states that 20 % of the population will earn 80 % of the total income in the population. According to Pareto this "law" should be universal and inevitable regardless of taxation, social and political conditions.

A general formulation would state that the fraction β of the population earns the fraction γ of the total income in the population. It shall be emphasized that β and γ do not in general add to one. Further, the elements in the population need not to be persons, and the measure for the individual elements could in principle be any well defined measure. The measure could also

be a measure of change for the individual elements in the population e.g. the time for decomposition of the elements. However, in this paper for convenience we will when referring to population refer to "persons" for the elements and to "income" for the measure. Since the statement of Pareto's law, the law has been investigated by a number of economists and other scientists resulting in a number of "refutations".

The Pareto distributions

In order to obtain an understanding of the complex of Pareto distributions, we start by introducing the classic Pareto distribution.

Consider again the basic equation

$N(x) = Ax^{-\alpha}$ (1)

It is obvious that $x \to \infty \Rightarrow N(x) \to 0$. However, since $x \to 0 \Rightarrow N(x) \to \infty$, the introduction of a smallest income k in the population is a necessity, i.e. N(k) =N. Thus, we have

$N_1(x) = Ax^{-\alpha}$; for $0 < k \leq x$ and 0 else; $= 0$ $A > 0$; $\alpha > 0$ (2)

Alternatively, we can define N(x) as

$N_2(x) = N$; $x = 0$ and $N_2(x) = Ax^{-\alpha}$ $0 < x$; $A > 0$; $\alpha > 0$ (3)

If the equation (2) is divided by N i.e. with the number of persons in the population we obtain $G_1(x) = N(x)/N = Ax^{-\alpha}/N$ and G(x) is the fraction of persons in the population having an income larger than x. The probability that a randomly selected person from the population has an income larger than x is thus $G_1(x)$. It is assumed that all persons in the population have the same probability of being selected.

$F_1(x) = 1 - G_1(x) = 1 - Ax^{-\alpha}/N$ is therefore a distribution function for the random variable X_1 which denotes the income for a randomly selected person from the population i.e.

$P(X_1 \leq x) = F_1(x)$ $(x \geq k)$.

The distribution of X_1 is denoted as the Pareto distribution of the first kind.

Since $F_1(k) = 1 - Ak^{-\alpha}/N = 0$, we have $k = (A/N)^{1/\alpha}$ or $A/N = k^{\alpha}$.

The modern notation for the Pareto distribution of the first kind is therefore

$$F_1(x) = 1 - \left(\frac{k}{x}\right)^{\alpha} \quad 0 < k \leq x \tag{4}$$

The density function for X_1 is

$$f_1(x) = \frac{\alpha k^{\alpha}}{x^{\alpha+1}} \tag{5a}$$

This is the Pareto distribution normally seen in most textbooks.

An alternative expression for $f_1(x)$ is

$$f_1(x) = cx^{-(\alpha+1)};\ c = \alpha k^{\alpha} \tag{5b}$$

The notation $X_1 \sim Par_1(k, \alpha)$ is normally applied.

In many situations the type 1 Pareto distribution works well. In some cases it is, however, necessary to have a distribution which is also defined for $x = 0$. This is obtained for the Pareto distribution of the second kind. This distribution is therefore naturally based directly on equation (3). The distribution function for the distribution is

$$F_2(x) = 1 - \left(\frac{c}{x+c}\right)^{\alpha} = 1 - \left(1+\frac{x}{c}\right)^{-\alpha};\ 0 \le x;\ 0 < c;\ 0 < \alpha \tag{6a}$$

The density function is:

$$f_2(x) = \frac{\alpha c^{\alpha}}{(x+c)^{\alpha+1}} \tag{6b}$$

The normal notation is $X_2 \sim Par_2(c,\alpha)$.

The distribution of X_2 is also denoted as the Lomax distribution (see Lomax; 1954).

It should be noted that when using the type 2 distribution it is in some fields applicable to introduce a location parameter μ where μ is real i.e. $\mu \in \mathbf{R}$.

This results in the following distribution function $F_{2a}(x)$ and density function $f_{2a}(x)$:

$$F_{2a}(x) = 1 - \left(\frac{c}{x+c-\mu}\right)^{\alpha} = 1 - \left(1+\frac{x-\mu}{c}\right)^{-\alpha} \tag{7a}$$

$$f_{2a}(x) = \frac{\alpha c^{\alpha}}{(x+c-\mu)^{\alpha+1}} \tag{7b}$$

The normal notation is $X_{2a} \sim Par_{2a}(\mu,c,\alpha)$. Sometimes, this distribution is denoted as the Pareto distribution of the second kind.

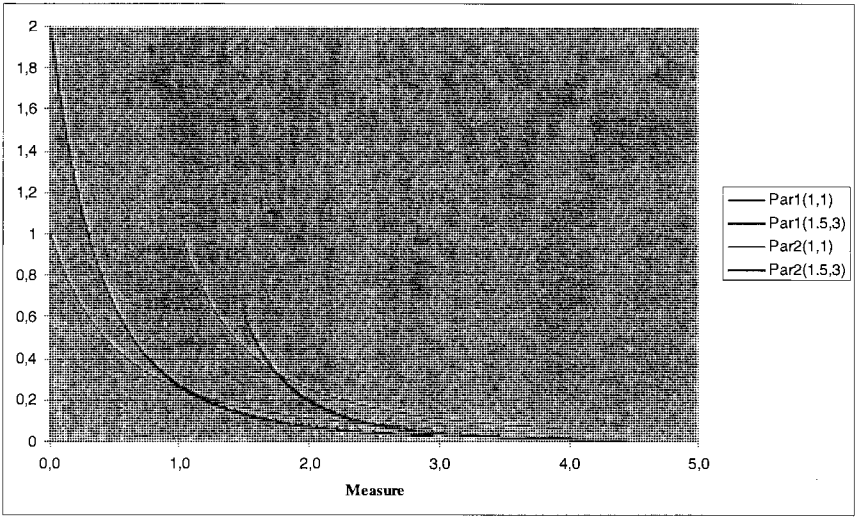


Figure 1. Density functions for the Pareto distribution of the first and the second kind, displaying the influence of the parameters. The notation is for the first kind $Par_1(k,\alpha)$ and for the second kind $Par_2(c,\alpha)$.

The moment No. r (r = 1, 2, · · ·) about zero for X_1 and X_2 is calculated in the usual way:

$$E(X_1^r) = \int_k^\infty x^r \frac{\alpha k^\alpha}{x^{\alpha+1}} dx = \frac{\alpha k^r}{\alpha - r}; \alpha > r \tag{8}$$

$$E(X_2^r) = \int_0^\infty x^r \frac{\alpha K_1}{(x+c)^{\alpha+1}} dx = \alpha K_1 c^{r-\alpha} B(\alpha - r, r+1) = \frac{\alpha c^r \Gamma(\alpha - r)\Gamma(r+1)}{\Gamma(\alpha+1)}$$

$$= \frac{c^r \Gamma(\alpha - r)\Gamma(r+1)}{\Gamma(\alpha)}; \alpha > r \tag{9}$$

where $B(\bullet,\bullet)$ denotes the beta function.

Due to the lack of symmetry of the density function the expected value for a Pareto distribution is not of great value for most applications. Therefore, in most cases the median x_m of the distribution will be of greater practical value than the expected value.

Other useful measures of location are the harmonic mean H and the geometric mean G where $H = \{E([X^{-1}]\}^{-1}$ or $H^{-1} = E([X^{-1}]$ and $\ln(G) = E(\ln(X)$.

For the Pareto distribution of the first kind the these measures of location together with the variance are given by Johnson et al. (1994, p. 577 and 578).

From equation (8) we obtain

$E(X_1) = \alpha k/(\alpha-1)$; $1 > \alpha$ and $Var(X_1) = \alpha k^2/[(\alpha-1)^2(\alpha-2)]$; $2 < \alpha$ (10a)

From equation (4) we obtain for the median: $F_1(x_{1m}) = 1-\left(\frac{k}{x_{1m}}\right)^{\alpha} = 0{,}5$ i.e.

$$x_{1m} = k2^{1/\alpha} \tag{10b}$$

$$H_1 = k(1+\alpha^{-1}) \tag{10c}$$

$$G_1 = ke^{1/\alpha} \tag{10d}$$

For the Pareto distribution of the second kind it is also without problems to calculate $E(X_2)$, $Var(X_2)$ and the median.

From the equations (9) we obtain:

$$E(X_2) = \frac{c}{1-\alpha} \text{ and } Var(X_2) = E(X_2^2) - (E(X_2))^2 = \frac{\alpha c^2}{(\alpha-1)^2(\alpha-2)} \tag{11a}$$

The median x_{2m} for X_2 is calculated from equation (6a):

$$F_1(x_{2m}) = 1-\left(\frac{c}{x_{2m}+c}\right)^{\alpha} = 0{,}5: x_{2m} = c(2^{1/\alpha}-1) \tag{11b}$$

However, the calculation of H_2 and G_2 is more complicated and the author has not found a reference which verifies the calculations given below.

$H_2^{-1} = E(1/X_2) = \int_0^{\infty} \frac{\alpha c^{\alpha}}{x(x+c)^{\alpha+1}} dx$. Using the substitution x = (c/t –c) we obtain the equation

$H_2^{-1} = \int_0^1 \frac{\alpha c^{\alpha} t^{\alpha}}{(1-t)} dt$ i.e. $H_2 = (\int_0^1 \frac{\alpha c^{\alpha} t^{\alpha}}{(1-t)} dt)^{-1}$.

According to Spiegel (1963, chap 12, problem 16) the integral $\int_0^1 \frac{\alpha c^{\alpha} t^{\alpha}}{(1-t)} dt$ diverges.

Therefore,

$$H_2 = 0 \tag{11c}$$

$$\ln(G_2) = E(\ln(X_2)) = \int_0^{\infty} \frac{\alpha c^{\alpha} \ln x}{(x+c)^{\alpha+1}} dx . \tag{11d}$$

This integral can only be evaluated using numerical methods, e.g. for $\alpha = 2$ and $c = 2$ we obtain $\ln(G_2) = -0{,}30685$ and thus $G_2 = 0{,}73576$.

Besides the Pareto distributions of the first and the second kind, both a third and a fourth kind of the distribution have been defined. These two distributions are both "natural" extensions of the distributions of the first and second kind and are denoted $Par_3(\mu, \sigma, \gamma)$ and $Par_4(\mu, \sigma, \gamma, \alpha)$, respectively. The distribution function $F_3(x)$ and $F_4(x)$ together with the density functions $f_3(x)$ and $f_4(x)$ for the two extended Pareto distributions are

$$F_3(x) = 1 - [1+\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}]^{-1};\ f_3(x) = \frac{\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma-1}}{\sigma\gamma[\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}+1]^2} \quad (12a)$$

$$F_4(x) = 1 - [1+\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}]^{-\alpha};\ f_4(x) = \frac{\alpha\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma-1}}{\sigma\gamma[\left(\frac{x-\mu}{\sigma}\right)^{1/\gamma}+1]^{\alpha+1}} \quad (12b)$$

The Pareto distribution of the fourth kind was defined by Arnold (1983). See also Arnold (2014).

It shall be emphasized that there exits a Pareto distribution which is based on the Pareto distribution of the second kind and also is denoted as the Pareto distribution of the third kind. According to Johnson et al (1994, p. 575) this Pareto distribution of the third kind was defined by Pareto himself and is denoted $Par_3P(c, \alpha, b)$. The distribution function $F_{3P}(x)$ and the density function $f_{3P}(x)$ of the distribution are, respectively

$$F_{3P}(x) = 1 - \frac{c^\alpha e^{-bx}}{(x+c)^\alpha} = 1-(1+\frac{x}{c})^{-\alpha}e^{-bx} \text{ and } f_{3P}(x) = \frac{c^\alpha e^{-bx}(b(x+c)+\alpha)}{(x+c)^{\alpha+1}} \quad (12c)$$

It is seen that the distribution function and the density function for the Pareto distribution of the second kind (equation (6a) and (6b)) can be obtained from equation (12c) by inserting b = 0. In equation (12c) the second expression for $F_{3P}(x)$ was given by Bottazzi (2007) and by Arnold (2014). Thus, it is obvious that the expression for $F_{3P}(x)$ given by Johnson et al. (1994, p. 575) is wrong ("C" should in the expression for $F_{3P}(x)$ have been "C^α").

So-called generalized Pareto distributions have been defined a number of times. Since the distribution of the inverse of a beta distribution is a Pareto distribution of the first kind is i.e. $Y \approx Be(p, 1) \Rightarrow X = 1/Y \sim Par_1(p, 1)$. This fact has inspired to the definition of some generalized Pareto distributions based on transformations of the inverse beta distribution. Feller (1971, p. 50) defined the distribution $X_F = 1/Y - 1$ where the distribution of Y is a beta distribution i.e.

$Y \sim Be(p,q)$. The density function of X_F is

$f_F(x) = \frac{x^{q-1}}{B(p,q)(1+x)^{p+q}}$ where B(p,q) is the beta function. The notion is $X_F \sim Par_F(p, q)$

The Feller distribution was further generalized (Arnold; 1983) defining the so-called Feller-Pareto distribution:

$X_g = \mu + \sigma(1/Y - 1)^{\gamma}$ where the distribution of Y is a beta distribution i.e. Y ~ Be(p,q). The notation for the FP distribution is X_g ~ FP(μ, σ, γ, p, q). The density function of X_g is:

$$f_g(x) = \frac{\Gamma(p+q)(\frac{x-\mu}{\sigma})^{q/\gamma-1}}{\gamma\sigma\Gamma(p)\Gamma(q)\left[(\frac{x-\mu}{\sigma})^{1/\gamma}+1\right]^{p+q}} = \frac{(\frac{x-\mu}{\sigma})^{q/\gamma-1}}{\gamma\sigma B(p,q)\left[(\frac{x-\mu}{\sigma})^{1/\gamma}+1\right]^{p+q}}$$

$(x > \mu; \in R; \sigma, \gamma > 0)$ (13)

The Pareto distributions of the first and second kind (equation (5a) and (6b) together with the special case of kind 2 (equation (7b) are special cases of the FP distribution i.e.

$FP(k, k, 1, \alpha, 1) = Par_1(k, \alpha)$; $FP(0, c, 1, \alpha, 1) = Par_2(c, \alpha)$ and

$FP(\mu, c, 1, \alpha, 1) = Par_{2a}(\mu, c, \alpha)$

The same is true for the distribution of X_F: $FP(0, 1, 1, p, q) = Par_F(p, q)$

Further, $FP(\mu, \sigma, \gamma, 1, 1) = Par_3(\mu, \sigma, \gamma)$ and $FP(\mu, \sigma, \gamma, \alpha, 1) = Par_4(\mu, \sigma, \gamma, \alpha)$

(see e.g. Arnold; 2014).

Choulakian & Stephens (2001) discussed the distribution X_{CS} which they denote as the generalized Pareto distribution. Respectively, the distribution function $F_{CS}(x)$ and the density function $f_{CS}(x)$ for X_{CS} are:

$F_{CS}(x) = 1 - (1 - k_1x/\alpha_1)^{1/k_1}$ and $f_{CS}(x) = (1/\alpha_1)(1 - k_1x/\alpha_1)^{(1-k_1)/k_1}$

$(0 \le x \text{ for } k_1 < 0;\ 0 \le x \le \alpha_1/k_1 \text{ for } k_1 > 0)$

The notation is $X_{CS} \sim Par_{CS}(k_1, \alpha_1)$ and it can be shown that for $k_1 < 0$ then

$FP(0, -\alpha_1/k_1, 1, -1/k_1, 1) = Par_{CS}(k_1, \alpha_1)$.

According to Johnson et al. (1994; p 614) the distribution of X_{CS} has also been discussed by other authors. The distribution has been applied in the analysis of extreme events especially in hydrology (see e.g. Choulakian & Stephens 2001).

Finally, with reference to Ljubo (1965) Johnson et al (1994; p. 614) defined a generalized Pareto distribution with the following distribution function ($\beta > 0$):

$$F_J(x) = 1 - \left(\frac{k+\alpha}{x+\alpha}\right)^{\alpha} e^{-\beta(x-k)} \text{ for } x \ge k \text{ and } F_J(x) = 0 \text{ for } x < k$$

For $k = 0$ and $\alpha = c$ this distribution function is identical to the distribution function $F_{3P}(x)$ given in equation (12c). It is, however, outside the scope of this paper to treat all these distributions further.

Though also the Pareto distribution of the second kind is a very interesting distribution, it is not possible in this paper to discuss this distribution in more detail. Therefore, in the rest of the paper only the Pareto distribution of the first kind is discussed together with the discrete Pareto distribution.

Some properties for the Pareto distribution of the first kind

Especially the Pareto distribution of the first kind has a number of interesting properties.

The distribution is invariant with respect to the transformation $Y = a X_1^r$, where a, r > 0.

$X_1 \sim Par(k,\alpha) \Rightarrow Y = a X_1^r \sim Par(ak^r,\alpha/r)$ (14a)

$$P(Y \le y) = P(a X_1^r \le y) = P(X_1 \le (\frac{y}{a})^{1/r}) = 1 - (\frac{k}{(y/a)^{1/r}})^{\alpha} = 1 - (\frac{ka^{1/r}}{y^{1/r}})^{\alpha} = 1 - (\frac{(k^r a)^{1/r}}{y^{1/r}})^{\alpha}$$

$$= 1 - (\frac{k'}{y})^{\alpha'}, \text{ where } k' = ak^r \text{ and } \alpha' = \alpha/r.$$

For $X_1 \sim Par(k,\alpha)$ consider the transformation $Y_{inv} = 1/X_1$.

$$F_{inv}(y) = P(Y_{inv} \le y) = P(1/X_1 \le y) = P(X_1 \ge 1/y) = 1 - F_1(1/y) = (\frac{k}{y^{-1}})^{\alpha} = (ky)^{\alpha} \ (0 \le y \le 1/k).$$

The density function of Y_{inv} is $f_{inv}(y) = \alpha k^{\alpha} y^{\alpha-1}$ $(0 \le y \le 1/k)$ (14b)

This is the distribution function for a scaled beta distribution with q = 1 and the scaling factor 1/k i.e. $X_1 \sim Par(k,\alpha) \Rightarrow Y = 1/X_1 \sim Y_B/k$ where $Y_B \sim Be(\alpha,1)$.

This distribution is called *the power-function distribution* (Johnson et al.; 1994; p. 607). However, within the power technology the distribution is also denoted as the Gates-Gauden-Schuhmann distribution and is applied to describe populations of sorted particles i.e. where no particle has a measure larger than 1/k.

The conditional distribution of X_1 is also a Pareto distribution of the first kind. If *a* and *b* are positive constants then

$$P(X_1 > a+b | X_1 > a) = \frac{P((X_1 > a+b) \cap (X_1 > a))}{P(X_1 > a)} = \frac{P(X_1 > a+b)}{P(X_1 > a)}$$

$$= \frac{(\frac{k}{a+b})^{\alpha}}{(\frac{k}{a})^{\alpha}} = (\frac{a}{a+b})^{\alpha} \qquad (14c)$$

Thus, $P(X_1 > a+b | X_1 > a) = P(Y_1 > a + b)$ where $Y_1 \sim Par_1(a, \alpha)$ i.e. $y \ge a$ while $x \ge k$.

Moment distributions for the Pareto distribution

When introducing the distributions of X_1 and X_2 it was assumed that the probability of selecting a person from the population was the same for all persons in the population. However, in a number of situations where each item in a population is characterized by a measure, it is applicable to consider a selection where each item is selected with a probability which is proportional to a function of the measure. The most simple situation is when the probability is proportional to the measure itself i.e. the probability of selecting an item with the size x is $\delta_1 x$ where δ_1 is a constant. In many fields the measure is referred to as the "size" of the item and is thus normally considered as a real number.

Let X denote the size (e.g. an income) for a randomly selected item (e.g. a person) where all items in the population have the same probability of being selected and let X_{p1} denote the size of an item selected randomly from the same population but with a probability which is proportional to the size of the item. The distribution function of X_{p1} is then

$F_{p1}(x) = \int_0^x \delta_1 t f(t) dt$ where f(t) is the density function for X.

Since $1 = \int_0^\infty \delta_1 t f(t) dt$ we have $\delta_1 = 1/E(X)$ and $E(X_{p1}) = E(X^2)/E(X)$

The distribution of X_{p1} is denoted as the first moment distribution of X.

This can be generalized to the case where the selection is proportional to the size in the r'th power. In this situation let the random variable X_{pr} denote the size of a randomly selected item. We then have $F_{pr}(x) = \int_0^x \delta_r t^r f(t) dt$ where $\delta_r = 1/E(X^r)$.

Thus, $E(X_{pr}) = E(X^{r+1})/E(X^r)$ which is the expected value for the items in the population weighted with respect to the size in the r'th power. Depending on the population and the measure characterizing the individual items, $E(X_{pr})$ is for different values of r often preferred to describe the population instead of E(X). This is of course true in econometrics but also in other fields e.g. in power technology where the items are particles and the measure may be the aerodynamic diameter of the particles in the population. The population may then be described by $E(X_{p3})$ which is the expected value of the particle diameters weighed with respect to the diameter in the third power i.e. also with respect to particle volume and particle mass. It should be emphasized that when the items are particles, the distribution of X is in many situations a lognormal distribution The distribution of X_{pr} is denoted as the r'th moment distribution of X and the distribution function is normally given as

$$F_{pr}(x) = \frac{1}{E(X^r)}\int_0^x t^r f(t)dt \qquad (15)$$

For some statistical distributions the moment distributions are of the same type as the "mother" distribution i.e. the distribution of X_{pr} is of the same type as the distribution of X but with different parameters. This is e.g. the case for the lognormal distribution and fortunately also for the Pareto distribution of the first kind.

Inserting the equations (5a) and (8) in equation (15) results in

$$F_{1pr}(x) = \frac{\alpha - r}{\alpha k^r}\int_k^x t^r \frac{\alpha k^\alpha}{t^{\alpha+1}}dt = \frac{\alpha - r}{k^{r-\alpha}}\int_k^x \frac{1}{t^{\alpha-r+1}}dt = \frac{\alpha - r}{k^{r-\alpha}}\left[\frac{t^{-\alpha+r}}{-\alpha+r}\right]_k^x$$

$$= -k^{\alpha-r}(x^{r-\alpha} - k^{r-\alpha}) = 1 - \left(\frac{k}{x}\right)^{\alpha-r} \qquad (15a)$$

Thus, the distribution of X_{1pr} is a Pareto distribution of the first kind i.e. $X_{1pr} \sim Par_1(k, \alpha\text{-}r)$

When the population consists of persons, and the measure is the income of the individual persons, X_{1pr} denotes the income for a person randomly selected with a probability proportional to the person's income. According to the equations (15a) and (10a), the population could then be described by $E(X_{1pr}) = (\alpha - 1)k/(\alpha - 2)$.

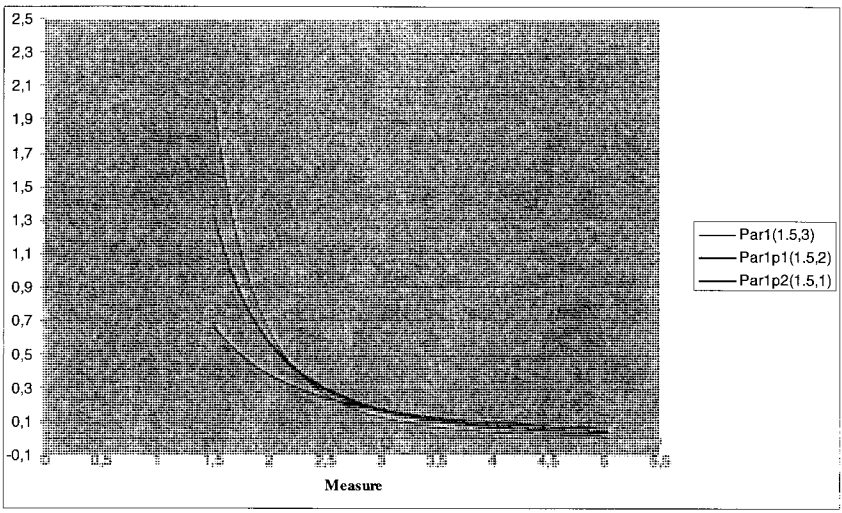


Figure 2. The density function for a Pareto distribution of the first kind with the parameters k = 1,5 and $\alpha = 3$ together with the two first corresponding moment distributions (i.e. r = 1 and r = 2) which also both are Pareto distributions of the first kind with the same k value but α is reduced by r (r = 1, 2).

The Lorenz function and the Gini coefficient

In applications where the individual items in a population are each characterized by a measure it is sometimes applicable to have a measure for the homogeneity of the items in the population. In this section we introduce the Lorenz function and the Gini coefficient which are used to estimate the homogeneity.

Consider again the population with N persons, where the income X_1 for a randomly selected person has a Pareto distribution of the first kind.

The number of persons in the population having an income less or equal to x is then $N_1(x) = NF_1(x)$.

The total income for the $N_1(x)$ persons earning less or equal to x is denoted $I_1(x)$.

The total income $dI_1(x)$ for the $dN_1(x)$ persons in the interval [x; x + dx] is then $dI_1(x) = xdN_1(x)$ i.e.

$dI_1(x) = x(dN_1(x)/dx)dx = xNf_1(x)\,dx$.

The total income for the number of persons having an income less or equal to x is thus

$I_1(x) = N\int_k^x tf_1(t)dt$ while the total income in the population is $I_1 = N\int_k^\infty tf_1(t)dt = NE(X_1)$ ($\alpha >$ 1) i.e. the total income for persons earning less or equal to x relative to the total income in the population is

$$R_I(x) = \frac{I_1(x)}{I_1} = \frac{I_1(x)}{NE(X_1)} = \int_k^x \frac{t}{E(X_1)} f_1(t)dt = 1-\left(\frac{k}{x}\right)^{\alpha-1} \qquad (16)$$

Thus, $R_I(x)$ is calculated from the distribution function for the first moment distribution of X_1 (see equation (15)). If $\alpha \to \infty$ then $R_I(x) \to 1$. This means that the income for all persons in the population is the same i.e. the income is k for all persons in A.

Since the $R_I(x)$ is in form of a distribution function, it could be asked which random variable is having this distribution function. The random variable corresponding to $R_I(x)$ is according to the remarks after equation (15) the distribution which is obtained, when the persons in the population are selected randomly with a probability proportional to the individual person's income.

The case where $\alpha \to \infty$ may in some situations be of interest as a reference. However, since $R_I(x)$ is not very useful for $\alpha \to \infty$ the function is rewritten and in stead of using the income x as the independent variable $y = F_1(x)$ is used. This results in the function:

$R_I(F_1(x)) = 1 - (1 - F_1(x))^{1-1/\alpha}$ or $R_I(y) = 1 - (1 - y)^{1-1/\alpha}$

This function is defined for $0 \le F_1(x) \le 1$ i.e. $0 \le y \le 1$ and the value space for the function is $0 \le R_I(F_1(x)) \le 1$ The function is denoted the Lorenz function and the graph of $R_I(F_1(x))$ is denoted the Lorenz curve. A special case of the Lorenz function is obtained for $\alpha \to \infty$ which results in $R_I(F_1(x)) \to F_1(x)$. Thus, the graph of the Lorenz function is in this case a straight line with a slope equal to one.

The Gini coefficient $G(\alpha)$ is a function of α and is defined as twice the area between the graph of the actual Lorenz function and the graph of the function $R_I(F_1(x)) = F_1(x)$ which graph is a straight line i.e. for the Pareto distribution of the first kind we have

$$G(\alpha) = 1 - 2\int_0^1 R_I(F_1(x))dF_1(x) = 1 - 2\int_0^1 R_I(y)dy = 1 - 2\int_0^1 (1-(1-y)^{1-1/\alpha})dy = \frac{1}{2\alpha - 1} \qquad (17)$$

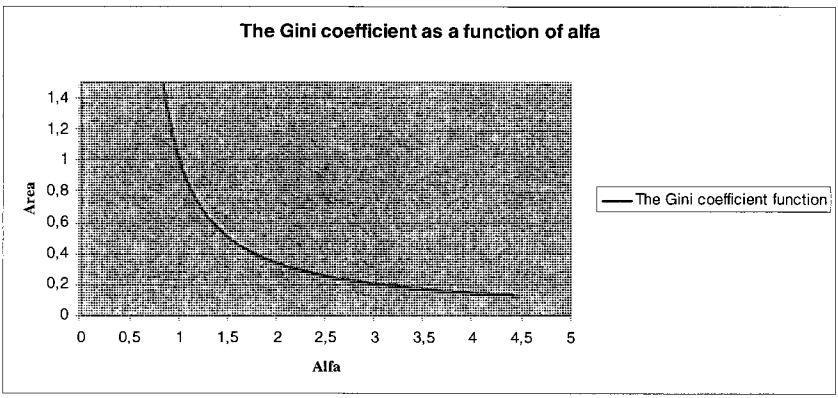


Figure 3. The Gini coefficient for a Pareto distribution of the first kind as a function of the parameter α.

The Gini coefficient is in fact a measure of homogeneity. The smaller the Gine coefficient will be the more homogeneous the population will be with respect to the measure defined.

It should be emphasized that neither the Lorenz function nor the Gini coefficient are unique to the Pareto distribution of the first kind. When the inverse function $x = F^{-1}(x)$ exists for a distribution function $y = F(x)$ of a random variable X the Lorenz function for the distribution is defined as

$$L(F(x)) = L(y) = \int_{x_m}^{F^{-1}(x)} tf(t)dt / E\{X\} \quad (0 \le y \le 1;\ 0 \le L(y) \le 1.$$

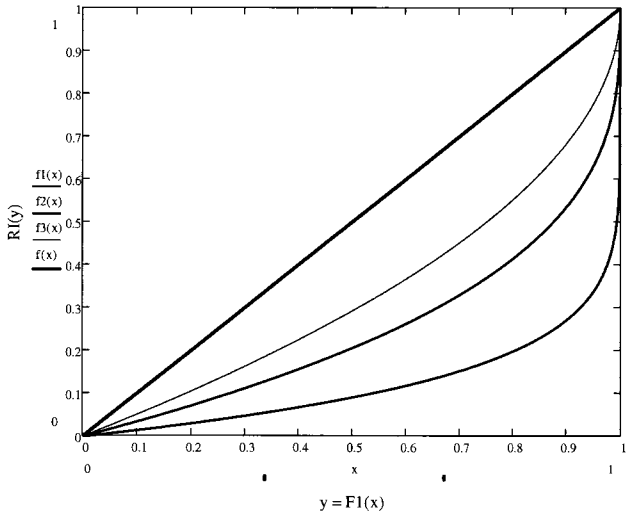


Figure 4. The graphs of three Lorenz functions for the Pareto distribution of the first kind.
The functions are denoted $f_1(x)$, $f_2(x)$ and $f_3(x)$ where $f_1(x)$ is in the lowest position. The corresponding values of α are $\alpha_1 = 1{,}16$ (Pareto's constant), $\alpha_2 = 1{,}5$ and $\alpha_3 = 2$. The figure displays the influence of α on the Gini coefficient which is twice the area between the graphs of the Lorenz functions and the straight line $f(x) = x$.

Thus, the Lorenz function can be calculated from the first moment distribution of X by inserting $x = F^{-1}(x)$ in the expression for the moment distribution. The Gini coefficient is in general defined as twice the area between the graph of the actual Lorenz function and the straight line $y = x$.

However, it can be expected that the calculations in general will be complicated.

The Pareto law

The Pareto Law was referred to in the Introduction. The law states that 80 % of the total income in the population was earned by 20 % of the persons in the population.

Using the notation given in the Introduction we have $\beta = 1 - F_1(x_0) = 0{,}20$ for $x = x_0$.

Then according to the Pareto Law we have $\gamma = 1\text{-}R_1(x_0) = 0{,}80$.

In this section we discuss the general ratio for the Pareto distribution of the first kind i.e.

$$r(x) = \frac{\beta}{\gamma} = \frac{1 - F_1(x)}{1 - R_1(x)} = \frac{\left(\frac{k}{x}\right)^{\alpha}}{\left(\frac{k}{x}\right)^{\alpha-1}} = \frac{k}{x} \qquad (18)$$

The ratio $r(x)$ is seen to be independent of α and we have $x = k/r(x)$.

For a given ratio $r(x_0) = r_0$ we therefore have $x_0 = k/r_0$ i.e. $r_0 = 1/4$ results in $x_0 = 4k$.

However, $\beta = 1 - F_1(x_0)$ and $\gamma = 1\text{-}R_1(x_0)$ are both functions of α.

Since $x = k/r(x)$ we have

$$\beta = 1 - F_I(x_0) = \left(\frac{k}{x_0}\right)^{\alpha} = \left(\frac{k}{k/r_0}\right)^{\alpha} = r_0^{\alpha} \quad \text{and } \gamma = 1 - R_I(x_0) = \left(\frac{k}{x_0}\right)^{\alpha-1} = r_0^{\alpha-1}.$$

E.g. for $\alpha = 2$ and $r_0 = 1/4$ we obtain $\beta = 1 - F_I(x_0) = 0{,}0625$ and $\gamma = 1\text{-}R_I(x_0) = 0{,}25$.

Note that β and γ do not add to one. 6,25 % of the population earn 25 % of the total income in the population.

The Pareto Law is only true when $\beta = (1/4)^{\alpha} = 0{,}20$ i.e. for $\alpha = \ln 5/\ln 4 = 1{,}16$ and therefore $\gamma = 1 - R_I(x_0) = (1/4)^{0{,}16} \approx 0{,}80$.

This results in the Gini coefficient $G(1{,}16) = 1/(2\times 1{,}16 - 1) = 0.758$.

The discrete Pareto distribution

Until now we have only treated continuous Pareto distributions defined on a population A where the individual items each was characterized by a measure. However, the items in A may also be characterized by being identical with respect to some well defined criteria. It is shown below, that this characterisation leads to the discrete Pareto distribution.

The discrete Pareto distribution is defined by the following probability function:

$$P(X = r) = cr^{-(\alpha+1)} \quad (r = 1, 2, \ldots, ; \alpha > 0) \tag{19}$$

Where $c = \left[\sum_{r=1}^{\infty} r^{-(\alpha+1)}\right]^{-1} = [\xi(\alpha+1)]^{-1}$ i.e. c is the inverse of Riemann's zeta function (see e.g. Spiegel; 1964, p.273). Therefore, the distribution of X is also denoted as the zeta distribution (Johnson & Kotz; 1969, p. 69). In the following it will be explained which deductions have resulted in this statistical function

Historically the zeta distribution is based on the so-called Zipt-Estroup law. This law states that the rth most frequent word in a (long) text has a frequency nr which is proportional to $1/r^{\rho}$ i.e. $n_r \propto 1/r^{\rho}$ or $n_r = c_1/r^{\rho}$ where c_1 is the constant of proportionality. In this equation r is called the frequency rank of a word. The most frequent word i.e. $r = 1$ has a frequency n_1 proportional to 1, the second most frequent word ($r = 2$) has a frequency n_2 proportional to $1/2^{\rho}$, the third most frequent word ($r = 3$) has a frequency n_3 proportional to $1/3^{\rho}$ and so on.

This may be generalized. Consider a population, A, of N items where some of the items are identical with respect to some well defined criteria. In this case there exists an intersection of A where subset No. r consists of the r most common items i.e. the number of items in the subset is n_r. Thus, subset No. 1 consists of n_1 identical items and this particular item is the

most common item in A. Let X map each of the subsets in the number of the number of the subset e.g. X maps subset No. 4 in 4. Thus, the value space of X is $\Omega_X = \{1, 2, 3, \ldots, M\}$.
When a randomly selected item from the population belongs to class No. r, the value of X is r. Therefore, the probability function of X is $P(X = r) = n_r/N$.
Let us now assume that n_r is proportional to $1/r^\rho$ i.e. $n_r \propto 1/r^\rho$ or $n_r = c_1/r^\rho$ where ρ is a positive constant. This results in $P(X = r) = n_r/N = c_1/(Nr^\rho)$.

Since we have $\sum_{r=1}^{M} \frac{c_1}{Nr^\rho} = 1$ we obtain

$c_1 = N[\sum_{r=1}^{M} \frac{1}{r^\rho}]^{-1}$ which results in the following probability function of X:

$P(X = r) = r^{-\rho}[\sum_{r=1}^{M} \frac{1}{r^\rho}]^{-1}$.

For large values of N and thus also of M it has in some applications been convenient to apply the approximation $[\sum_{r=1}^{M} \frac{1}{r^\rho}]^{-1} \approx [\xi(\rho)]^{-1}$ i.e. the inverse of Riemann's zeta function. Therefore we now write

$P(X = r) = r^{-\rho}[\xi(\rho)]^{-1}$ $(r = 1, 2, \ldots; \rho > 1)$. (20)

The expected value E(X) and the variance Var(X) can be shown to be

$E(X) = \xi(\rho\text{-}1)/\xi(\rho)$ and $Var(X) = \xi(\rho\text{-}2)/\xi(\rho) - (\xi(\rho\text{-}1)/\xi(\rho))^2$

(See https://www.randomservices.org/random/special/Zeta.html)

In case A consists of the words from a very long text, according to the Zipt-Estroup law the value of ρ is then slightly larger than 1 and the distribution should only in this case be denoted as the Zipt-Estroup distribution or only as the Zipt distribution.
In order to obtain a resemblance to the Pareto distribution of the first kind given in equation (5b) ρ is sometimes replaced by $\rho = \alpha + 1$ (see equation (19)).
It is outside the purpose to discuss the applications of the zeta distribution. However, in a situation with a population where some of the items are identical with respect to some well defined concept, it should be considered whether the zeta distribution could be applied to describe the population.

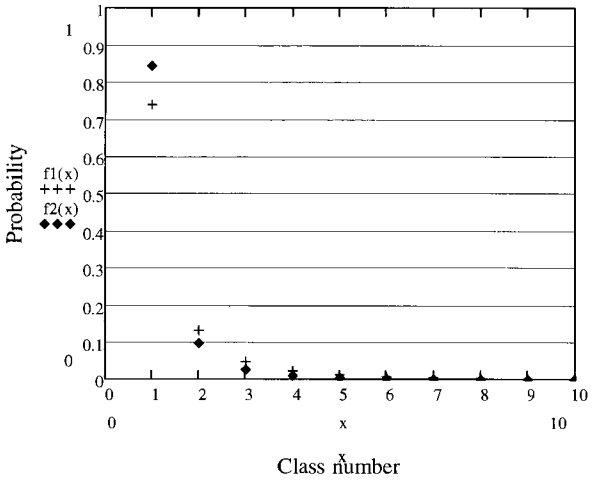


Figure 5. The probability functions for two zeta distributions i.e. for two discrete Pareto functions. The parameters for $f_1(x) = P(X_1 = x)$ are $c_1 = 0{,}7400$ and $\rho_1 = 2{,}4749$, respectively. The f2(x) probability function has the parameters $c_2 = 0{,}8450$ and $\rho_2 = 3{,}1$.

Conclusion

As mentioned in the introduction a large amount of literature on the Pareto distribution and the application of the distribution exists.

In this paper an overview is presented illustrating basic concepts of the distribution including the so-called generalized Pareto distributions. Some of the applications have been referred to. It has been shown how the Pareto distribution of the first kind (equation (5a)) and the corresponding discrete Pareto distribution (equation (20)) can both be deduced from a population of items under relevant assumptions.

In the population from which the "classic" Pareto distribution is deduced the individual items are all characterized by a measure. This situation leads to the definition of a random variable X which denotes the measure for a randomly selected item from the population where all items have the same probability of being selected. Without assuming any distribution for X this situation leads to the definition of the moment distributions for X. Therefore, also to the function denoted the Lorenz function is defined together with a Gini coefficient which is a measure for homogeneity of the items in the population in relation to the measure defined. In this paper only the case where the distribution of X is a Pareto distribution of the first kind has been discussed. It was shown that the moment distributions of X_1 were also Pareto distributions of the first kind. In situations with other distributions than the Pareto distribution the calculations may be very complicated and only numerical methods should be applied. The

Pareto law was generalized and the dependence of α emphasized. The "classic" 80-20 statement and the corresponding Gini coefficient were confirmed as $\alpha = 1{,}16$ and $G(1{,}16) = 0{,}758$ respectively.

The discussion of the zeta distribution and its applications has due to lack of space only been on the basic level. According to the author's opinion the zeta distribution is a very interesting distribution and a number of new applications for the distribution may be expected in the future.

Acknowledgement

The author is grateful to Professor Ph.D. Helle Rootzen, DTU Compute, Technical University of Denmark, and to MSc Morten Lindholm, ALK-Abelló, Hørsholm, Denmark for their valuable comments and suggestions during the development of this paper.

References

Arnold, B., C. (1983). Pareto Distributions. (Fairland MD: International Publishing House).

Arnold, B., C (2014). Univariate and multivariate Pareto models. Journal of Statistical Distributions and Applications *1:11*, 1-16. (Open Access paper)

Botazzi, G. On the Pareto Type III Distribution. Laboratory of Economics and Management, Sant' Anna School of Advanced Studies, Pisa, Italy. (On the type 3 distribution defined by Vilfredo Pareto)

Choulakian, V., and Stephens, M., A (2001). Goodness-of-Fit Tests for the Generalized Pareto Distribution. Technometrics *43*, 478-479.

Feller, W. (1971). An Introduction to Probability Theory and its Applications, 2nd ed., vol. 2. (New York: John Wiley & Sons).

Johnson, N., L, and Kotz, S. (1969). Discrete Distributions (Distributions in statistics 1). (New York: John Wiley & Sons).

Ljubo, M. (1965). Curves and concentration indices for certain generalized Pareto distributions. Statistical Review *15*, 257-260. (In Serbo-Croatian, English summary)

Lomax, K., S (1954). Business failures: Another example of the analysis of failure data. Jour. Amer. Stat. Assoc. *49*, 847-852.

Workload and empathy in general practice

Justin A. Charles[1,2], Peder Ahnfeldt-Mollerup[2], Jens Søndergaard[2], Troels Kristensen[2,3]

[1]Center for Medical Humanities, Compassionate Care, and Bioethics, Stony Brook University, Stony Brook, New York
[2]Research Unit of General Practice, Department of Public Health, University of Southern Denmark, J.B. Winsløws Vej 9, DK-5000 Odense C, Denmark
[3]Danish Center for Health Economics (DaCHE)

Work in progress, do not refer to or cite without permission from the authors.

Background:
Physician empathy is important for optimal physician and patient outcomes. Some hypothesized that there is an association between high workload and low level of empathy. Therefore, we aim to analyze the association between measures of workload and physician empathy among GPs.

Method:
A web-based survey comprising GP demographic, professional and workload characteristics, as well as the Jefferson Scale of Empathy (JSE-HP), was distributed to a random, stratified sample of 1,196 Danish GPs. Workload characteristics included activity measures, workload inside and outside of the clinic, as well as perceived workload-burden (PWB) measures of selected socioeconomic patient types. By means of regression models we analyzed the association between GP workload and empathy. Factor analysis and weighted averages were performed to reduce the number of covariates for PWB and create rounded, global measures of perceived workload.

Results:
A total of 464 GPs (38.8%) in 406 clinics completed the questionnaire. There was no link between empathy and activity measures of workload. An "U-shaped" relationship emerged between empathy and proxies for overall GP perceived workload burden, such that high empathy scores occurred at both high and low levels of PWB. These findings were consistent across model types. A high PWB associated with "Difficult Communication" patient types was associated with higher empathy levels.

Conclusion:
GPs who perceive a higher workload and GPs who perceive a lower workload appear to be more emphatic than the other GPs.

1. Introduction

Physician empathy is well-documented as an important aspect of patient care.[1] Previous studies have demonstrated that higher levels of physician empathy are associated with improved patient satisfaction, adherence, and health outcomes, as well as improved physician well-being and decreased burnout.[2-5] While some have argued that empathy is an innate and unchangeable characteristic, there is increasing evidence that physician empathy can be influenced by aspects of the clinical environment, such as workload.[6,7] General Practitioners (GPs) report that a heavy workload serves as a barrier to empathy.[8] Incorporating empathy in patient care requires time and effort to actively listen, understand patient's concerns, and communicate compassionately. GPs already have many responsibilities to manage during relatively short visit times, including managing multiple chronic conditions, performing health maintenance and screening, and counseling and educating patients.[9] Workload in general practice continues to increase, resulting in GPs needing to see more patients in less time.[10,11] These shorter consultations have been associated with reduced empathy in clinical encounters, while longer consultations are linked to increased empathy.[12,13]

However, time spent on patient encounters in the clinic is just one contributor to the larger, multifactorial concept of workload in general practice.[14] Some patient visits are done out-of-office, via modalities such as telephone triage and house calls.[15] In addition, a high proportion of patient care involves charting in the electronic health records (EHR) and communicating with patients and other healthcare providers.[16] There are also more structural determinants of GP workload. One determinant is the organization and efficiency of a GP's practice and utilization of non-physician healthcare providers.[17,18] Another, is issues related to the healthcare system, such as increasing administrative duties and task shifting between primary and secondary care.[19]

The complexity of a GP's patient panel, in terms of psychosocial and multimorbidity-related characteristics, is another important factor in GPs workload.[20] For instance, patients from deprived areas require more frequent consultations and provide communication challenges to physicians due to issues surrounding language, culture, and health literacy.[21,22] While being empathic is important to managing the complex physical, mental, and social health needs of these patients, GPs who practice in more deprived areas appear to have high levels of workload and job stress, and lower levels of empathy.[23,24]

This finding aligns with GP perception and appears to suggest that there is a link between high workload and low physician empathy. As a result, we hypothesize that higher measures of workload will be associated with lower empathy levels in our study. However, the relationship between empathy and workload is not well understood. This may be due to the complex nature of the two concepts.[25,26] Therefore, we aim to examine

the association between GP workload and empathy by using a survey containing a set of workload measures that reflect multiple aspects of workload.

2. Methods

2.1 Setting

Danish GPs are independent contractors for the regional health authorities. and serve as the gatekeepers of the Danish Healthcare System.[27] Nearly every Danish citizen is listed with a specific GP, who offers comprehensive primary care to patients of all ages, and is seen for initial non-emergent consultations prior to specialist referral in most cases. This organization results in high levels of continuity of care in Danish General Practice.[28] A total of 1600 patients are listed with a GP and they consult the GP on average, slightly less than 7 times per year.[29]

2.2 Sample

The sample of 1195 GPs was selected from a list of 2926 email addresses of GPs obtained in 2013 from the Danish General Practitioners' Union (PLO). Several studies indicate that GP characteristics and workload may vary across different practice types and locations.[30,31] Therefore, to improve the representativeness of the study participants, a "stratified proportion allocation" was applied to a random sample of GPs based upon combinations of practice type (partnership, non-partnership) and practice location (urban, rural, mixed). This process is described at length in our previous paper.[32]

2.3 Survey

A web-based survey was sent to GPs currently practicing in Denmark in December 2016. The survey was comprised of questions about GP demographic (gender and age), professional (practice type, practice location, and job satisfaction), and workload characteristics, as well as the Jefferson Scale of Empathy (JSE-HP). The workload characteristics included: number of GP full-time equivalents (FTE) in a practice; average number of hours worked per week in clinic and out of clinic; average allocated consultation time; average number of patients seen per day; presence and type of employment outside of clinic. For measures that required average values, GPs were asked to estimate based on a normal work day or week.

Additionally, perceived workload-burden (PWB) of a set of patient types was assessed. This was done by asking 11 questions about how strongly the following patient types contributed to PWB using a 0-9 Likert scale: children (younger than 5); foreign-born; elderly (above 65-years-old); living alone; single parent; unemployed; low education; high mobility; early pension; unemployment benefits; outside the labor market (in job center courses). 8 of these 11 patient types were originally included in the Care Need Index (CNI), a well-established measure of workload based on socioeconomic patient types used in Sweden.[33]

The three new characteristics reflect patients receiving early pension, unemployment benefits, or job training outside the labor market. Patients are entitled to early pension because of permanent inability to work, frequently secondary to a physical or mental health issues. Unemployment and job loss are often associated with poor physical and mental health.[34] Patients with these three characteristics are offered comprehensive socioeconomic services in Denmark, which require administrative duties that often fall on the GP.[35] Combined, the administrative and health-related issues in these patient populations may result in more frequent, complicated visits, representing a higher workload for physicians. Indeed, there are reports of an association between patients utilizing social services and increased physician workload.[21,36]

The Jefferson Scale of Empathy Health Professions version (JSE-HP) was administered to measure physician empathy.[37] The JSE-HP is designed to measure empathic behavior orientation specifically among health professionals, such as physicians. The creators of the scale operationalize empathy with the following definition: "Empathy is a predominantly *cognitive* (rather than an affective or emotional) attribute that involves an *understanding* (rather than feeling) of experiences, concerns and perspectives of the patient, combined with a capacity to *communicate* this understanding, and an *intention to help*."[37] Each item features a statement using a 7-point Likert scale (with 7 indicating strong agreement), such as "I try to imagine myself in my patients' shoes when providing care to them." The total score ranges from 20-140, with a higher overall score indicating higher levels of empathy. The scale has been extensively validated in terms of its convergent, discriminant, concurrent, and predictive validity, as well as internal consistency, test-retest reliability, and low social desirability bias.[1,38,39]

2.4 Data Collection

In December 2016, our survey was emailed to 1195 Danish GPs using SurveyXact software, accompanied by a cover letter describing the study. Participants were offered the equivalent of $20 for their time as per protocol of the Danish Multipractice Committee, who approved our study. Two reminder emails were sent to non-respondents 2 and 4 weeks after the initial email. This study received ethical approval from the Institutional Review Board at Stony Brook University. Data was uploaded to Statistics Denmark and all identifiers were encrypted to allow for anonymous analysis of the data.

2.5 Statistical Analysis

2.5.1 Descriptive Statistics

Descriptive statistics were used to describe GP demographic, clinic, and workload characteristics. Dummy variables were used to code for responses of binary nature, such as gender and practice. Numerical responses to questions were treated as continuous variables.

2.5.2 Regression Model

Both OLS and Random Effect (RE) models were used to analyze the association between the empathy scores and sets of covariates. RE models account for inter-group variation at the clinic level, as the random sample of data were drawn from a population of individual GPs nested in clinics. The RE model takes the following form:

$$JSE_{ij} = \alpha + x_{ij}\beta_1 + u_j + v_{ij} \qquad (1)$$

where JSE represents the continuous empathy score for GP_i in GP clinic j. X_{ij} is a row of explanatory variables and uj is the unobserved GP clinic specific effect referring to the conditional mean of the JSE score in clinic j. The model assumes that the u_j is independent of all explanatory variables. The intraclass correlation coefficient (rho) was calculated to determine the proportion of the total variance ($\sigma_u + \sigma_v$) that is accounted for by clustering within clinics (σ_u).

The OLS model ignores the intra-group error correlation and takes the form:

$$JSE_j = \alpha + x_i\beta_1 + v_i \qquad (2)$$

Different sets of covariates were used for each estimation procedure for both OLS and RE models. For all models, a common set of demographics (age and gender), professional characteristics (practice location, practice type, job satisfaction), and workload measures were included. Alternative models also included: 1) high PWB; 2) overall PWB; 3) workload factors; as mentioned below. The weighted and unweighted versions of these three covariates were incorporated into OLS and RE regression models. Next, we inspected for linear relationships between each covariate and empathy, and tested for the relevance of non-linear transformation of parameters. Based on this, the square functional form of overall PWB was also included as a covariate in the models containing the linear form of overall PWB.

2.5.3 Reduction of Patient Type Covariates

The number of patient type covariates were reduced to offer global measures of workload and diminish the potential effect of multicollinearity in regression analysis. This was accomplished in two ways: 1) Overall workload indicator; 2) factor analysis.

1) Overall perceived workload burden (PWB) indicator

Two overall PWB indicators were constructed: One was an unweighted indicator, where all patients count the same across all patient types. The second was a weighted indicator, based on a weighted average formula (3):

$$\bar{x} = \sum_{i=A}^{K} w_i' x_i \qquad (3)$$

where the individual GP's assessment of the patient type is x_i and the normalized weight for patient type i ($w_i^{'}$) was calculated based on the original weights (w_i) from Table 1 in the following way:

$$w_i^{'} = \frac{w_i}{\sum_{j=A}^{K} w_j} \qquad (4)$$

The expression (3) yields the weighted PWB indicator where some patient types (such as foreign-born) contribute more to the weighted mean than other patients with a lower weight. These indicator variables were used to create two dummy proxy variables entitled "high PWB" and "weighted high PWB", which represent GPs above the 90th percentile for unweighted and weighted overall PWB respectively.

2) Factor Analysis

Second, exploratory factor analysis (EFA) was conducted to reduce the number of covariates into potential underlying constructs of workload factors. Bartlett test of sphericity (BTS) and Kaiser-Miller-Olkin measure of sampling adequacy (MSA) were used to test for appropriateness of factor analysis. Next, principal-component factors method using orthogonal, varimax rotation was applied. Internal consistency was assessed for each factor and for the whole scale using Cronbach alpha. These factors were constructed and prepared for analysis by creating an unweighted and weighted measure of each factor. The weighted version involved multiplying the variables in the factor by the factor loading of each variable divided by the sum of all factor loadings.

3. Results

3.1 Study Sample

A total of 464 GPs in 406 clinics completed the survey for a response rate of 38.8%. Representativeness-ratios for each stratum in the analysis showed good representativeness in terms of age, practice type, and practice location, with minimal over- or under-representativeness. These results can be found in detail in our previous paper[32]

3.2 Descriptive Statistics

The average empathy score among this population was 117.85, compared to a maximum score of 140. The descriptive statistics for GP demographic and professional characteristics demonstrate a slight male predominance (53.4%) and an average age of 54.9 years. Most GPs work in partnership practices (72.0%) and in an urban location (49.3%). 79.7% of GPs were somewhat or very satisfied with their job. A more detailed account of these measures can be found in our previous paper.[32]

The descriptive statistics for the GPs' self-reported workload characteristics in Table 1 shows an average of 2.76 GP full-time equivalents (FTE) within a range from 1 (p5) to 6 (p95). GPs worked an average of 40.68 hours per week in clinic, ranging from 30 (p5) to 55 (p95) hours, and 5.86 hours per week outside of clinic, ranging from 1 (p5) to 20

(p95) hours. The allocated consultation time is reported to be between 10 min (p5) and 15 min (p95) with an average of 14.33 minutes. Furthermore, the reported number of visits per day is between 18 (p5) and 34 (p95), with an average of about 24 patients a day.

Table 1: Self-reported workload characteristics among Danish GPs

Workload characteristic	Mean	CV	p5	median	p95
# GP FTE capacities per clinic	2.76	0.56	1	2	6
Avg. hours working in clinic per week±	40.68	0.20	30	40	55
Avg. hours working outside clinic per week±	5.86	1.60	1	4	20
Avg. allocated consultation time (min.)±	14.33	0.14	10	15	15
Avg. # patients per day±	23.91	0.21	18	24	34
Type of employment outside clinic					
A. None	30.8%				
B. Out of hours' services	45.5%				
C. Consultant work	25.2%				
D. Teaching	16.6%				
E. Research	0.03%				
F. Professional policy activity	13.1%				
G. Other	11.2%				
Contribution of patient types to GP PWB (Scale 0-9)					
A. Foreign-born	6.82 (11.8%)	0.28	3	7	9
B. High mobility	6.76 (11.7%)	0.32	2	7	9
C. Outside labor market	6.39 (11.0%)	0.31	2	7	9
D. Elderly (65 and over)	6.18 (10.7%)	0.31	2	7	9
E. Public transfer payment	5.55 (9.6%)	0.39	2	6	8
F. Unemployed (ages 18-65)	4.96 (8.6%)	0.43	1	5	8
G. Early pension	4.93 (8.5%)	0.49	1	5	8
H. Low education	4.62 (8.0%)	0.47	1	5	8
I. Single parent	4.24 (7.3%)	0.51	0	5	7
J. Living alone	3.81 (6.6%)	0.59	0	4	7
K. Children (under age of 5)	3.59 (6.2%)	0.66	0	4	7
Workload indicators					
Overall PWB	5.26	0.29	2.36	5.50	7.27
Weighted Overall PWB‡	5.49	0.27	2.56	5.73	7.49
High PWB (N= 55)	7.33	0.06	6.90	7.18	8.27
Weighted High PWB‡ (N = 47)	7.58	0.05	7.13	7.49	8.48
"Economically Disadvantaged"	4.76	0.35	1.75	5.13	7.00
Weighted "Economically Disadvantaged" ‡	6.59	0.23	6.67	7.00	8.33
"Difficult Communication"	4.77	0.36	1.66	5.11	7.09
Weighted "Difficult Communication" ‡	6.64	0.23	3.69	6.97	8.51

CV: Coefficient of variation (SD/Mean), FTE: Full time equivalents; PWB: Perceived workload-burden
±GPs estimated based on normal day or week. ‡Weighted by contribution of each patient type

Most GPs were employed outside of clinic (69.2%). They were involved in out of hours' or call work (45.5%), consultant work (25.2%), teaching (16.6%), research (0.03% and professional policy activity (13.1%), as well as other activities that were not specified. The average PWB for the sample of GPs ranged from a maximum average score of 6.82 (out of 9) for foreign-born patients compared to a minimum average score of 3.59 for children under the age of 5. The second most burdensome patient type was high mobility patients (those who switch practice frequently) (6.76), followed by patients outside the labor market (6.39) and elderly patients (6.18).

3.3 Reduction of Patient Type Covariates

3.3.1 Overall PWB

The average "overall PWB" indicator was 5.26, and became 5.49 when each of the 11 patient types (A-K) was weighted by its contribution to PWB. These weights are listed in parentheses next to the mean score for each patient type. There was large variation in the overall PWB across GPs, both weighted and unweighted. However, the weighted PWB resulted in generally higher values. The weighted scores ranged from 2.56 (p5) to 7.49 (p95) reflecting a three-fold difference.

3.3.2 Factor analysis

The BTS ($\chi 2$ (55) = 2973.84, $p < 0.00001$) and MSA test (0.909) for appropriateness of factor analysis for the 11 patient type variables indicated that factor analysis is appropriate and relevant. Cronbach alpha for the whole scale was 0.91 indicating high internal consistency. Three factors were retained with eigenvalues of 5.76 and 1.11, and 1.01 prior to rotation, accounting for 52%, 10%, and 9% of the total variance respectively. However, since factor 3 contained only one variable, indicating its instability, the factor analysis was conducted again to retain only two factors.

The first factor, entitled "Economically Disadvantaged" contained the following 8 items: low education; unemployed; single parent; children; living alone; public transfer payment; early pension; outside labor market. Factor coefficients were all above 0.60, except for children, which was 0.36, with an average of 0.72. Cronbach alpha for this factor was measured to be 0.90. The second factor, entitled "Difficult Communication" contained the following 3 items: foreign-born; high mobility; elderly. Factor coefficients were at least 0.50, with an average of 0.68. Cronbach alpha for this factor was measured to be 0.63.

3.4 Regression Model Results

Overall, the models in Table 2 show that the alternative sets of covariates explained variation in the empathy score among GPs ($p<0,0001$ for F-tests). The adjusted r^2 values show that the included variables explained between about 5% and 10% of the variation in empathy scores. The models using the alternative sets of workload measures: a) High

PWB dummy; b) Overall PWB & PWB^2; c) workload factors; all reveal a significant association between the included workload measures and empathy.

Table 2: Association between empathy score and workload characteristics among GPs

Characteristic	RE 1	OLS 1	OLS 2	OLS 3
Gender	-1.41	-1.417	-1.196	-1.689
Age	0.011	0.0111	0.0149	0.00511
Non-partnership practice (ref. partnership)	-1.661	-1.662	-1.938	-1.689
Urban clinic (ref. mixed)	-1.065	-1.072	-0.837	-0.826
Rural clinic (ref. mixed)	0.418	0.412	0.332	0.688
Somewhat or Very Satisfied with job	2.889*	2.897*	3.122*	2.718*
GP FTE capacities per clinic	-0.696	-0.699	-0.675	-0.654
Avg. hours working in clinic per week±	0.0698	0.0691	0.0852	0.0842
Avg. hours working outside clinic per week±	0.0354	0.0359	0.0446	0.0225
Avg. allocated consultation time (min.)±	-0.191	-0.187	-0.187	-0.0918
Avg. # patients per day±	-0.0767	-0.0748	-0.0557	-0.0534
Type of employment outside clinic				
A. No employment outside clinic	1.186	1.191	1.359	1.815
B. Out of hours' services	-1.889	-1.900	-1.743	-1.665
C. Consultant work	-0.39	-0.411	-0.488	-0.324
D. Teaching	4.480**	4.480***	4.514***	4.353**
E. Research	2.406	2.484	2.333	2.433
F. Professional policy activity	3.764*	3.753*	3.620*	4.356**
G. Other	2.567	2.552	2.413	2.599
Workload Indicators				
Weighted High PWB‡	5.383**	5.383**		
Weighted Overall PWB‡			-6.120***	
Weighted Overall PWB^2 ‡			0.671***	
Weighted "Economically Disadvantaged" ‡				-0. 0402
Weighted "Difficult Communication" ‡				0.446*
Constant	118.4***	118.3***	129.0***	111.5***
N (GP respondents)	464	464	464	464
Number of clinics (groups)	406			
σ_u	1.51			
σ_v	10.58			
Rho	0.02			
r^2 overall	0.1			
r2		0.1	0.117	0.0906
r^2 adjusted		0.0616	0.0775	0.0496
F		2.873	3.612	2.248
P	<0.0001	<0.0001	<0.0001	0.00166
Akaike Information Criterion (AIC)		3523.9	3517	3530.8

Chi^2	52.79			
Variance Inflation Factor (VIF)	1.41	1.41	3.72	1.44

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; ‡Weighted by contribution of each patient type
±GPs estimated based on normal day or week; PWB: Perceived workload-burden; RE: Random Effect

The dichotomous variable high PWB in models RE1 and OLS1 shows that a high perceived workload, above the 90th percentile, is linked to higher empathy. The squared nonlinear functional form in OLS2 shows a quadratic relationship (U-shape) with high empathy levels present at both low and high workloads. Based on the parabolic estimation function, the minimum empathy score is 115.05, which occurs at a PWB of 4.56. OLS3 demonstrates that the Difficult Communication factor is linked to higher empathy.

The statistical significant covariates among the independent variables had positive signs and were stable across the models. GPs who were somewhat or very satisfied with their job had higher empathy scores. There was no relationship between empathy and gender, age, practice type, practice location, measures of time spent in or out of clinic, allocated consultation time, number of patients seen per day, or number of GP FTE in a clinic. GPs who participated in teaching or political advocacy work outside of clinic had higher empathy scores than their peers.

RE and OLS models were nearly identical for all proposed estimations as demonstrated by the comparison of the two models RE1 and OLS1. Intraclass correlation coefficients (rho) in the RE models ranged from 0 to 0.02, showing that observations within clinics were no more similar than observations from different clinics. For all workload variables that were weighted, unweighted versions reached the same level of significance in each model, so only weighted versions were used. As evidence by the VIF <10, there was no evidence of multicollinearity. However, the squared functional form resulted in a higher VIF as expected.

4. Discussion

This study contributes to the literature in several ways. First, we investigated the relationship between GP empathy and workload, in terms of volume and frequency of work inside and outside the clinic, as well as perceived burden of patients with selected socioeconomic characteristics. Second, we included three new patient characteristics that are relevant to examine perceived GP workload, but have not been used previously in general practice. Third, we collected and constructed updated weights determined by GPs relating to the perceived workload-burden of 11 different patient types. Fourth, we explored the effect of using weighted averages and factor analysis to develop indicators of overall perceived workload. Finally, we explored the functional form of the link between these overall workload measures and empathy. Sensitivity analysis was undertaken with respect to alternative sets of covariates to examine for consistency across different regression models.

We found that the empathy score was higher both at low and high PWB. This demonstrates a non-linear and more complex relationship between workload and empathy than we and others hypothesized. On one hand, GPs with high PWB from deprived patients have been demonstrated to have lower empathy.[24] This is consistent with GP reports that workload pressure from complex patients serves as a barrier to empathic care.[8] On the other hand, GPs who are more empathic may self-select to practice in more underserved areas, where there is often a higher workload.[21] Subsequently, maintaining empathy requires effort, self-awareness, and self-regulation, and therefore, it may be associated with a higher workload [40] GPs who found patients in the "Difficult Communication" factor to be more burdensome to workload were more empathic. GPs who spend time and effort attempting to communicate in an empathic, patient-centered way may have higher levels of stress and compassion fatigue, which contributes to higher perceived workload.[21] Further studies should aim to identify proposed mechanisms and causal models of the association between these two complex concepts.

There was no significant relationship between empathy and measures of workload related to time spent in and out of clinic, allocated consultation time, or number of patients seen per day. This is consistent with two other studies that show no link between empathy and workload in terms of schedule density, current and global workload.[41,42] Conversely, Neumann *et al.* found that perceived busyness of physicians was associated with lower empathy levels.[43] Perhaps the lack of association in this study reflects a relatively lower workload among Danish GPs, compared to the UK, for instance, where workload pressure seems to have a large impact on physician performance.[10,11,44] However, this may not be an accurate assumption, as in 2016, 58% of Danish GPs were bothered by the fact that their work had become more demanding, compared to 25% in 2012.[29]

GPs involved in teaching and political advocacy work were more empathetic than those employed in other activities or not at all. Teaching and policy work specifically require strong interpersonal skills and GPs who become involved in these activities may have a predisposition towards empathy. GPs who believe interpersonal factors contribute to job satisfaction have been demonstrated to have higher empathy levels than their peers.[32]

The representative sample of GPs provided information regarding their PWB of important socioeconomic patient types. GPs viewed foreign-born patients as contributing the most to their workload burden. Foreign-born patients are typically given a high weight on measures of deprivation that reflect workload, such as the Jarman Index and the Danish Deprivation Index (DADI).[21,45] Previous literature indicates that these patients consult GPs at higher rates than native-born populations and often require translator services and assistance in navigating the healthcare system, which lead to increased workload.[44,46,47]

The survey used in this study has strengths and limitations. It allowed for the assessment of empathy scores and PWB by patient types using validated scales.[33] These measures allow for analysis of these latent variables by using indicators. Because the study population reflected a large sample size (13.5% of all Danish GPs), which was representative of the larger population, the data gathered provides an important perspective on the empathy levels and PWB of Danish GPs. However, this survey also had its limitations. GPs had to estimate their average workload, which may not have been fully accurate. Physicians have been reported to overestimate the amount of time spent during patient visits on self-report measures.[11] Combining the survey with more objective assessments of workload, such as EHR records or national databases, may have provided a more accurate depiction of the variables measured.[48]

Another limitation of this study is that there were several elements of workload that were not included in our model. List size and time spent on administrative activities, such as using the EHR were not included.[14,49] Additionally, we did not measure task shifting to primary care and number of non-physician healthcare providers.[19] Finally, the study only allowed for a cross-sectional analysis of workload and empathy, and cannot attribute causality of their relationship, nor demonstrate changes in either with time.

Using factor analysis allowed for an understanding of the underlying structure and reduction of the 11 patient types to two factors, entitled "Difficult Communication" and "Economically Disadvantaged." The model appeared sufficient as all factor loadings were above 0.30 and each factor retained had at least 3 items.[50] However, three of the items in the analysis were cross-loaded onto two factors, meaning that they loaded with coefficients of ≥ 0.32 on more than one factor. These items were retained because removing items may have compromised the integrity of the data set. Additional models should be attempted that apply oblique rotation, which allows factors to correlate, and may eliminate cross-loading issues. This theoretical model should be tested with confirmatory factor analysis to verify this preliminary work.

Moreover, indicator variables were created for overall and high PWB based on the ranking GP's gave to each patient type. These covariates were used to test models of the relationship between empathy and workload. An overall indicator of PWB is desirable as it provides one centralized measure of workload and avoids conflicting individual components.[51] Further studies using these measures of PWB can help to verify their use in approximating workload.

The JSE-HP empathy score is treated as a continuous, dependent variable in this study. This is in contrast to our previous study, where we dichotomized empathy scores by splitting respondents at a score of 120.[32] Using a dichotomized dependent variable reduces the effect of random errors in the measurement, erroneous self-report responses,

and distribution characteristics (e.g., skewed). Dichotomization also allows for a simpler interpretation of the link between covariates and empathy score in terms of odds ratios. However, there are disadvantages to dichotomizing originally continuous variables.[52] The empathy score derived from the JSE-HP does not have a naturally categorical structure, nor any theoretically meaningful cutoff points.[37] Dichotomization also reduces the observed variation as it does not allow for members in the same category to have different scores. Therefore, in line with De Coster *et al.,* we find that it is more appropriate to treat empathy score as a continuous variable. [52]

Random effect models were used in addition to OLS models to account for the possibility of regression errors correlated within GP clinics. However, the 464 GPs were nested in a relatively large number of clinics (406), indicating that clustering may have minimal effect. This was verified by calculating the intraclass correlation coefficient, which ranged from 0 to 0.02, showing a negligible effect of clustering on the variance of the data. The empathy scores within clinics are no more similar than observation from different clinics. This suggests that using random effect does not add robustness to the analysis, and therefore, OLS models were primarily used.

5. Conclusion

The overall workload indicators created can contribute to better measurement of GP perceived workload in future research. GPs who perceive a higher workload appear to be empathic.

References

1. Hojat M, Gonnella JS, Nasca TJ, Mangione S, Vergare M, Magee M. Physician empathy: Definition, measurement, and relationship to gender and specialty. *The American journal of psychiatry.* 2002;159.
2. Canale SD, Louis DZ, Maio V, et al. The relationship between physician empathy and disease complications: An empirical study of primary care physicians and their diabetic patients in Parma, Italy. *Academic Medicine.* 2012;87(9):1243-1249.
3. Hojat M, Louis DZ, Markham F, Wender R, Rabinowitz C, Gonnella JS. Physicians' Empathy and Clinical Outcomes for Diabetic Patients. *Academic medicine : journal of the Association of American Medical Colleges.* 2011;86.
4. Kim SS, Kaplowitz S, Johnston MV. The Effects of Physician Empathy on Patient Satisfaction and Compliance. *Evaluation & the Health Professions.* 2004;27(3):237-251.
5. Krasner MS, Epstein RM, Beckman H, et al. Association of an educational program in mindful communication with burnout, empathy, and attitudes among primary care physicians. *Jama.* 2009;302(12):1284-1293.
6. Caspi A, Silva PA. Temperamental Qualities at Age Three Predict Personality Traits in Young Adulthood: Lougitudinal Evidence from a Birth Cohort. *Child Development.* 1995;66(2):486-498.
7. Neumann M, Edelhauser F, Tauschel D, et al. Empathy decline and its reasons: a systematic review of studies with medical students and residents. *Academic medicine : journal of the Association of American Medical Colleges.* 2011;86(8):996-1009.
8. Derksen FAWM, Hartman TCo, Bensing JM, Lagro-Janssen ALM. Managing barriers to empathy in the clinical encounter: a qualitative interview study with GPs. *The British Journal of General Practice.* 2016;66(653):e887-e895.
9. Fiscella K, Epstein RM. So much to do, so little time: care for the socially disadvantaged and the 15-minute visit. *Archives of internal medicine.* 2008;168(17):1843-1852.
10. Thompson M, Walter F. Increases in general practice workload in England. *The Lancet.* 2016;387(10035):2270-2272.
11. Hobbs FDR, Bankhead C, Mukhtar T, et al. Clinical workload in UK primary care: a retrospective analysis of 100 million consultations in England, 2007-14. *Lancet (London, England).* 2016;387(10035):2323-2330.
12. Bayne H, Neukrug E, Hays D, Britton B. A comprehensive model for optimizing empathy in person-centered care. *Patient education and counseling.* 2013;93(2):209-215.
13. Lelorain S, Sultan S, Zenasni F, et al. Empathic concern and professional characteristics associated with clinical empathy in French general practitioners. *European Journal of General Practice.* 2013;19(1):23-28.
14. van den Berg MJ. *Workload in general practice.* Utrecht, NIVEL; 2010.
15. Huibers L, Moth G, Andersen M, et al. Consumption in out-of-hours health care: Danes double Dutch? *Scandinavian Journal of Primary Health Care.* 2014;32(1):44-50.
16. Chen MA, Hollenberg JP, Michelen W, Peterson JC, Casalino LP. Patient care outside of office visits: a primary care physician time study. *J Gen Intern Med.* 2011;26(1):58-63.
17. Wensing M, van den Hombergh P, Akkermans R, van Doremalen J, Grol R. Physician workload in primary care: what is the optimal size of practices? A cross-sectional study. *Health Policy.* 2006;77(3):260-267.
18. Morgan PA, Shah ND, Kaufman JS, Albanese MA. Impact of physician assistant care on office visit resource use in the United States. *Health services research.* 2008;43(5 Pt 2):1906-1922.
19. Baird B, Charles A, Honeyman M, Maguire D, Das P. *Understanding pressures in general practice.* King's Fund London; 2016.
20. Osborn R, Moulds D, Schneider EC, Doty MM, Squires D, Sarnak DO. Primary Care Physicians In Ten Countries Report Challenges Caring For Patients With Complex Health Needs. *Health Affairs.* 2015;34(12):2104-2112.
21. Pedersen AF, Vedsted P. Understanding the inverse care law: a register and survey-based study of patient deprivation and burnout in general practice. *International Journal for Equity in Health.* 2014;13:121.
22. Balarajan R, Yuen P, Machin D. Deprivation and general practitioner workload. *British Medical Journal.* 1992;304(6826):529-534.
23. Mercer SW, Jani BD, Maxwell M, Wong SY, Watt GC. Patient enablement requires physician empathy: a cross-sectional study of general practice consultations in areas of high and low socioeconomic deprivation in Scotland. *BMC Fam Pract.* 2012;13:6.
24. Mercer SW, Higgins M, Bikker AM, et al. General Practitioners' Empathy and Health Outcomes: A Prospective Observational Study of Consultations in Areas of High and Low Deprivation. *Annals of family medicine.* 2016;14(2):117-124.
25. Robieux L, Karsenti L, Pocard M, Flahault C. Let's talk about empathy! *Patient education and counseling.* 2018;101(1):59-66.
26. Michtalik HJ, Pronovost PJ, Marsteller JA, Spetz J, Brotman DJ. Developing a model for attending physician workload and outcomes. *JAMA Internal Medicine.* 2013;173(11):1026-1028.

27. Nørøxe KB, Pedersen AF, Bro F, Vedsted P. Mental well-being and job satisfaction among general practitioners: a nationwide cross-sectional survey in Denmark. *BMC Family Practice.* 2018;19:130.
28. Pavlic DR, Sever M, Klemenc-Ketis Z, et al. Strength of primary care service delivery: a comparative study of European countries, Australia, New Zealand, and Canada. *Primary health care research & development.* 2018;19(3):277-287.
29. Organisation of General Practitioners (PLO), Denmark: Fact sheet on Danish General Practice 2017. https://www.laeger.dk/sites/default/files/plo_faktaark_2017.pdf. Accessed January 2, 2019.
30. Lin H-C, Xirasagar S, Laditka JN. Patient perceptions of service quality in group versus solo practice clinics. *International Journal for Quality in Health Care.* 2004;16(6):437-445.
31. McGrail MR, Humphreys JS, Joyce CM, Scott A, Kalb G. How do rural GPs' workloads and work activities differ with community size compared with metropolitan practice? *Australian journal of primary health.* 2012;18(3):228-233.
32. Charles J, Ahnfeldt-Mollerup P, Søndergaard J, Kristensen T. Empathy Variation in General Practice: A Survey among General Practitioners in Denmark. *International Journal of Environmental Research and Public Health.* 2018;15(3):433.
33. Sundquist K, Malmstrom M, Johansson SE, Sundquist J. Care Need Index, a useful tool for the distribution of primary health care resources. *Journal of epidemiology and community health.* 2003;57(5):347-352.
34. Brand JE. The Far-Reaching Impact of Job Loss and Unemployment. *Annual review of sociology.* 2015;41:359-375.
35. Bardhan A, Jaffee D, Kroll C, Daemmrich A, Bredgaard T. The Welfare State as an Investment StrategyDenmark's Flexicurity Policies. In: Oxford University Press; 2013.
36. Muldoon L, Rayner J, Dahrouge S. Patient poverty and workload in primary care: study of prescription drug benefit recipients in community health centres. *Canadian family physician Medecin de famille canadien.* 2013;59(4):384-390.
37. Hojat M. *Empathy in health professions education and patient care.* Cham: Springer; 2016.
38. Hojat M, Mangione S, Nasca TJ, et al. The Jefferson scale of physician empathy: Development and Preliminary psychometric data. *Educ Psychol Meas.* 2001;61.
39. Hojat M, Zuckerman M, Magee M, et al. Empathy in medical students as related to specialty interest, personality, and perceptions of mother and father. *Personality and Individual Differences.* 2005;39(7):1205-1215.
40. Gleichgerrcht E, Decety J. Empathy in Clinical Practice: How Individual Dispositions, Gender, and Experience Moderate Empathic Concern, Burnout, and Emotional Distress in Physicians. *PLoS ONE.* 2013;8(4):e61526.
41. DiMatteo MR, Hays RD, Prince LM. Relationship of physicians' nonverbal communication skill to patient satisfaction, appointment noncompliance, and physician workload. *Health Psychology.* 1986;5(6):581-594.
42. Zantinge EM, Verhaak PF, de Bakker DH, Kerssens JJ, van der Meer K, Bensing JM. The workload of general practitioners does not affect their awareness of patients' psychological problems. *Patient education and counseling.* 2007;67(1-2):93-99.
43. Neumann M, Wirtz M, Bollschweiler E, et al. Determinants and patient-reported long-term outcomes of physician empathy in oncology: a structural equation modelling approach. *Patient education and counseling.* 2007;69(1-3):63-75.
44. Croxson CHD, Ashdown HF, Hobbs FDR. GPs' perceptions of workload in England: a qualitative interview study. *The British Journal of General Practice.* 2017;67(655):e138-e147.
45. Jarman B. Underprivileged areas: validation and distribution of scores. *British medical journal (Clinical research ed).* 1984;289(6458):1587-1592.
46. Balarajan R, Yuen P, Soni Raleigh V. Ethnic differences in general practitioner consultations. *Bmj.* 1989;299(6705):958-960.
47. Burt J, Lloyd C, Campbell J, Roland M, Abel G. Variations in GP–patient communication by ethnicity, age, and gender: evidence from a national primary care patient survey. *British Journal of General Practice.* 2016;66(642):e47-e52.
48. Arndt BG, Beasley JW, Watkinson MD, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Annals of family medicine.* 2017;15(5):419-426.
49. Young RA, Burge SK, Kumar KA, Wilson JM, Ortiz DF. A Time-Motion Study of Primary Care Physicians' Work in the Electronic Health Record Era. *Fam Med.* 2018;50(2):91-99.
50. Osborne J, Costello AB. Best Practices in Exploratory Factor Analysis: Four Recommendations for Getting the Most from Your Analysis. *Practical Assessment Research & Evaluation.* 2005;10(7).
51. Smith PC, Street A. Measuring the Efficiency of Public Services: The Limits of Analysis. *Journal of the Royal Statistical Society Series A (Statistics in Society).* 2005;168(2):401-417.
52. DeCoster J, Iselin A-MR, Gallucci M. A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods.* 2009;14(4):349-366.

Is the Prevention Paradox Useful in Suicide Prevention Theory and Strategies?

Mogens Nygaard Christoffersen, senior researcher emeritus
VIVE – The Danish Center for Social Science Research

Abstract

Little progress has been made in the areas of suicide prediction, explanation, and prevention mainly because of the difficulty inherent in examining a low base-rate behavior such as suicidal behavior. This study assesses the utility of the prevention paradox in improving suicide prevention theory and hence prevention strategies. This paradox would claim that the majority of cases with an adverse outcome (suicidal behaviour) derive from low- to moderate-risk groups, and only a few from a high-risk group. A number of variables based on suicide prevention theories are collected from administrative registers, and used to identify a high-risk group within a population of six birth cohorts (N = 300,000) whose members had the greatest likelihood of executing suicidal behavior in the period from birth to adulthood. The paradox holds true, and gives some empirical evidence for suicide prediction, explanation, and prevention that will serve to inform policymaking and treatment. Family background and family child-rearing factors such as parental mental illness, suicidal behavior, parental violence, unemployment and family separation all contributed independent information to prediction of suicide attempts. The young person's hospitalization for psychiatric disorder (ADHD, anxiety, depression, severe stress reaction and adjustment disorders), and being a victim of violence or a sex offence contributed to the explanation model. The results clearly indicate that the method to predict suicidal behavior is much better than chance ($p<0.0001$), but it seems impossible to identify a minor group (<12%) in the population from whom the suicide attempts are exclusively recruited. On the other hand, two thirds of the observed suicide attempts were included in the high-risk group. In the high-risk group eight out of ten turned out to be false-positive and one third of the observed suicide attempts derived from the low- to moderate-risk groups. Our example illustrates how the applicability of Rose's prevention strategy allows for the possible complementarity of population and targeted measures.

Background and significance

Suicidal behaviour is a multifactorial problem where individuals exposed to multiple risk factors are at the highest risk for suicide attempts and completed suicide (Christiansen et al., 2013; Qin & Nordentoft, 2005; Qin, 2003). Increased knowledge of risk factors is encouraged so that efforts targeting at-risk groups can be reinforced or improved. The selective prevention paradigm is based on the assumption that it is possible to identify a small group within the population (say, 5 to 10 per cent) at high-risk for suicidal behaviour from whom suicide or

attempted suicide are exclusive recruited. This approach contrasts with the prevention paradox. Geoffrey Rose's prevention paradox attains when the majority of cases with an adverse outcome come from a population of low or moderate-risk, and only a few from a 'high-risk' group (Rose, 1992; Rose, 1993). The present study intends to test whether the prevention paradox applies to the initiation of suicidal behavior, as recorded in longitudinal administrative data from Denmark.

Much attention has been focused on identifying potential victims of suicidal behaviour (Maris et al., 1992; Large et al., 2011; Christiansen et al., 2013; Evans et al., 2004). However, little headway has been made in recent decades in the areas of prediction, explanation, and prevention (Van Egmond & Diekstra, 1990) because of limitations imposed by (a) the low base rate of suicidal behaviour, and (b) the impossibility of ever demonstrating scientifically that any single intervention has been effective in preventing suicidal behaviour (Goldney, 2000). No single intervention has been shown in a well-documented randomized control trial to reduce suicide. Only a few studies have met randomized controlled trial research criteria; even when results from similar trials are synthesized using meta-analytical techniques, numbers are insufficient to detect differences (Hawton et al., 1998) Mann et al., 2005; Zalsman et al., 2016).

The first step in the proposed study is to identify potential causes of suicidal behaviour based on suicidal behaviour prevention theories. We will determine whether these well-known risk factors when combined in a life events analysis are highly correlated with first-time suicidal behaviour in a nationwide population. Second, we will evaluate whether most cases of adverse outcomes (suicide attempts) derive from combined low- and moderate-risk populations, and a relative few from high-risk groups.

The present study

The definition of suicide attempts includes behaviour that conforms to the following three conditions: (i) Suicide attempts that leads to hospitalization, (ii) assessment of the trauma as an act of self-mutilation according to the international statistical classification of injuries when discharged from hospital, (iii) the trauma is included in a specified list of traumas traditionally connected with suicide attempts: cutting in wrist (carpus), firearm wounds, hanging, self-poisoning with drugs, pesticide, cleaning fluids, alcohol or carbon monoxide. Intentional self-harm according to hospitals admissions in a psychiatric ward are also included (Christoffersen et al., 2003). This definition is comprehensive because the criteria, such as self-harm, do not differentiate between non-suicidal self-injury (NSSI) and behavior with a suicidal intent (Christoffersen et al., 2015).

Analysis

Potential predictors are identified from the preventive theories (Christoffersen et al., 2003; Christoffersen et al., 2007). The first step consists of selecting pertinent risk factors[1] based on theory.

Next, risk factors are sorted according to predictive value and capacity for explaining suicidal behaviour. We test whether the following risk factors are precursors of suicidal behaviour in the nationwide population study (see definitions in Appendix A):

- 1.a. Contemporary situation and opportunities: ADHD is a proxy for risk taking and impulsiveness.
- 1.b. Local community, social control, values and norms: living in a disadvantaged area, ethnically segregated housing
- 2.a. Developmental theories, disadvantages during the formative years: disadvantaged parenting (e.g., parental substance abuse), parental violence, parental unemployment, family separation, and the child in public care
- 2.b. Decision-making processes in high-risk groups that explain differences in suicidal behaviour in the context of present and future resources and conditions: education underachievement, victim of violence, psychiatric disorder (anxiety, depression, PTSD), substance abuse (i.e., drug addiction, alcohol abuse), and somatic disadvantages
- Theories of genetic and biological factors: family history of suicidal behaviour or mental health disorders (included in section 2.a., Table 1).

Statistical model

In order to maximize predictive efficiency, the estimation of parameters is based on the maximum likelihood method via the discrete-time Cox regression model. This model is used to allow for changing covariates over time (Allison, 1982).

Each individual is observed until time t, at which point either an event occurs or the observation is censored, by reaching the age limit, because of death, or the individual is lost to observation for other reasons. Individuals were excluded from the case group and controls after the first event. The person-years at risk were constructed the age groups 15 to 29 for the total birth cohorts born 1980 to 1985 who were living in Denmark at the age of 15 years. This makes up to 300,000 individuals and 4 million person-years.

When the discrete time unit is a calendar year, it is difficult to use continuous-time methods. Problems arise when the time intervals are large (e.g., one calendar year) so that more than one individual experiences an event in the same

[1] Risk is as a probability for an event (or outcome) within a specified population. In the present study the outcome is first time suicide attempt. The outcome is binary – either it occurs or it does not occur. We then define a risk factor as a correlate that is shown to precede the outcome of interest according to Kraemer and colleagues (Kraemer et al., 2005).

time interval (Allison, 1982). A discrete-time model is more appropriate for the estimation of parameters as it treats each individual history as a set of independent observations. It has been shown that the maximum likelihood estimator can be obtained by treating all the time units for all individuals as though they were independent, when studying first-time events (Allison, 1982).

The estimated hazards of first-time suicide attempts within the following year were estimated by following equations:

$$\pi(Y = 1) = \frac{e^{\beta_0+\beta_1+\beta_{2+\ldots}}}{1+e^{\beta_0+\beta_1+\beta_{2+\ldots}}} \quad (1)$$

weights[2] are estimated by the Greek letter β, and e is a constant, which also is the base of the natural logarithm. These estimations are done using the whole database and compared to actual observed suicide attempts based on hospital archives.

For each person and each calendar year, e.g. 1996, the first time hazards are calculated and named: π_{1996}. The hazards for not having a suicide attempt in year 1996, given that the person has not had an earlier attempt is therefore: $(1 - \pi_{1996})$.The hazard for having a suicide attempts at least once during the follow-up is one minus the hazards of not having an attempt any of the years (Christoffersen & Joshi, 2015):

$$1 - (1 - \pi_{1996})(1 - \pi_{1997})(1 - \pi_{1997}) \ldots (1 - \pi_{2014}). \quad (2)$$

This will be an estimate of the hazard of reaching the 29th birthday and a least once having attempted suicide according to hospitals files[3]. We then compare the observations with the hazard for having attempted suicide at least once, and calculate the false positive, the false negative and other measures such as sensitivity[4], specificity[5], etc. (Woodward, 2000).

Measuring predictive accuracy

A central question is how to assess the accuracy of predictions of suicidal behaviour (Christoffersen & Joshi, 2015). Various criteria for assessing accuracy have been developed in order to answer the question and new problems have arisen. One question could be: Is our method to predict suicidal behaviour better

[2] The weights (or parameters) are estimated according to 'maximum-likelihood'-method which gives the best possible prediction based on the most informative risk factors among the available factors in a step-wise selection process. The step-wise selection method has been criticized (Greenland, 1989) and it has been argued that the selection of parameters should be epidemiological rather than statistical, but the step-wise selection method seems appropriate and less time-consuming under the present circumstances.

[3] How do we estimate the succeeding hazards for young persons who actually have one incidence? We could imagine that nothing had happened and then estimate the persons' hazards for each of the succeeding years. Instead we estimate the hazard of the succeeding year(s) to be zero because we want an unbiased estimate of the first-time incidents.

[4] Sensitivity is the numbers of observed and expected in relation to number of observed; see Table 3: (a/(a+d).

[5] Specificity is the number of observed and expected persons, who did not attempt suicide in relation to observed persons not attempted; see Table 3: b/(b+c).

than chance? Here we face the problem that if only a small part of the persons is suicidal (low base rate of, for instance, 1 %) a researcher who predicts 'no suicidal behaviour' for everybody will be correct most of the times while a researcher who by chance predicts 'no suicidal behaviour' in, for instance, 20 % of the presented cases will be correct much less often. Special analytical tools (Receiver Operating Characteristic, ROC or Relative Improvement over Chance, RIOC) have been developed to capture if a screening method is better than chance. The ROC method is unaffected by base rates and unaffected by researcher's biases for or against Type I (false negative) or Type II (false positive) errors[6] and the method can be used to choose between screening instruments (Loeber & Dishion, 1983; Mossman, 1994; Mossman, 2000).

However, there is no single method that can answer completely the question of a prediction device's accuracy (Woodward, 1999). Even if we have a screening method that is clearly better than chance we still face problems of false positives. The problem is especially in focus when the predictions are used in an individual decision about suicide prevention measures. We have therefore chosen to include an additional aspect in measuring predictive efficiency: the proportion of 'false positives', taking advantage of the fact that methods for estimating predictive accuracy are well developed for dichotomous predictors and outcome (Blanc, 1998).

The false positive rate is commonly defined as the probability that a person who was not suicidal during the follow-up period was predicted as suicidal. The number of false positives is usually related to the total number of persons who were actually not suicidal (Woodward, 1999). Instead we prefer to focus at the number of false positives in relation to the total number of predicted suicidal behaviour during the follow-up period because this information can be decisive for a decision about individual suicide preventing measures. We would like to know the answer to the following question: how many of the predicted children will actually engage in suicidal behaviour within the follow-up period?[7] In the present paper we call this fraction 'the false positive percentage'.

Besides 'false positive', and 'false negative', 'sensitivity' and 'specificity' will be estimated, but the proportion of false positive turns out to be especially crucial when choosing between preventive methods targeting high-risk group of potential offenders.

Finally, we estimate the number of observed first-time attempters in the high-risk group and compare it with the observed numbers derived from the low- to moderate risk-group.

[6] The false negative is the wrong prediction of 'first-time suicide attempt' in cases where suicide attempt is observed, subsequently. The false positive is the wrong prediction of 'first-time suicide attempt' when no suicide attempt is observed in the follow-up period.

[7] According to Douglas Mossman and others we are misusing the term 'false positive rate' (Mossman, 2000; Woodward, 1999).

Data: Administrative registers

Administrative registers are useful because reporting biases are reduced when a series of administrative registers are linked together on the basis of personal identity numbers. Data are not aggregated or clustered but individually collected, independently. The collected data has at least four attributes of particular interest: (1) In contrast to new modern concepts such as 'big data' using data sources as Facebook, Google, and Twitter (Provost & Fawcett, 2013) professional agencies decide to incorporate data into the files on the basis of established criteria and manualised decisions. (2) Data are registered prospectively, that is information gathered in calendar year 't+1' has no influence whatsoever on data filed in calendar year 't'. The register data is not subject to back-filling. Even if later ('t+1') acquired information reveals errors these errors are not corrected for the year 't'. (3) Data are collected independently from various numbers of agencies. In other words, agency A has no knowledge of any register data completed by agency B. Finally, (4) data completely cover all the calendar years from their birth until early adulthood.

Personal identity numbers are initially used to link information for each individual from birth together with information about their parents. Later, the personal identity numbers are substituted with an encrypted number for security and ethical reasons.

Some of the previous nationwide population studies included only few of the risk factors and they all missed some of the most potential risk factors such as parental or offspring's substance abuse, parental violence, the child in public care, severe physical condition (Christiansen et al., 2013; Christiansen, 2007; Jablonska et al., 2009a; Hjern & Allebeck, 2002), except a few studies of suicide attempts and completed suicides (Christoffersen et al., 2003; Christoffersen et al., 2007).

We determine whether the above risk factors are found in any year before the first suicide attempt. Some risk factors have a high relative risk, e.g. the period after discharge from a psychiatric hospital (Large et al., 2011; Qin & Nordentoft, 2005), but only a few individuals are exposed. Other risk factors, on the other hand, are quite common but have a significant although low relative risk. We could estimate the attributable-fractions that express reduction in incidence of suicidal behaviour that would have been achieved if the population had not been exposed at all, compared to the empirical exposure pattern. On the basis of each individual's timeline of events, we calculate the hazards for attempting suicide from age 15 through 29 years and estimate the true positive rate (sensitivity) and specificity. We test whether, for example, the small percentage of the population with the estimated highest risk represents the majority of observed persons with suicidal behaviour. We have demonstrated the method in a previous study predicting criminal behaviour (Christoffersen & Joshi, 2015).

Results

Incidence rates of first-time suicide attempt with the given comprehensive definition of suicide attempts were between 290 and 370 per 100,000 individuals in age group 16 to 29 (see Figure 1). Lifetime estimate of first-time suicide attempt among 16 to 29 years old was 4.5 per cents of the population born 1980 to 1985.

Potential causes of suicidal behaviour are identified based on prevention theories. Four types of theories of prevention of suicidal behaviour were identified and their derived risk factors were located. Nineteen risk factors from all four types of theories of prevention were associated with an increased risk of first-time suicide attempt (see Table 2). Table 2 shows the adjusted discrete time Cox-model, the covariates, percentage of person-years exposed to respective risk factors, estimate of covariates, and Standard errors, estimate of Odds ratio. Nine of the nineteen risk factors had an effect size of more than 1.5 (odds ratio). Four of these had an effect size more than 2, when adjusted for other risk factors and age.

Decision-making processes in high-risk group, and developmental theories about disadvantages during the formative years seemed to make up the most influential factors. Certain high-risk groups have a significant higher risk of attempting suicide: Young people with no vocational training (odds ratio: 2.11), or not graduated from high school (odds ratio: 1.65), adolescents with somatic handicap (odds ratio: 1.40). In the high-risk group we also found adolescents with alcohol abuse (odds ratio: 2.42) or drug abuse (odds ratio: 1.75). Victims of a violent crime or sex offence (odds ratio: 1.55) were among the most at risk group together with adolescents with depression (odds ratio: 1.95), severe stress and adjustment disorder e.g. PTSD (odds ratio: 1.99).

Developmental theories and disadvantages during the formative years could constitute a part of the explanation of these relatively high risks such as parental mental illness (odds ratio: 1.26), parental suicidal behaviour (odds ratio: 1.15), parental violence (odds ratio: 1.17), family separation (odds ratio: 1.21), parental long-term unemployment (odds ratio: 1.12), and children in care (odds ratio: 1.29).

Theories of local community, social control, values and norms included indicators such as living in a disadvantaged area. The years living in these areas seemed to increase the risk of attempted suicide (odds ratio: 1.30) while being a non-Danish citizen were associated with relatively high risk (odds ratio: 1.13). Actual indicators of changing contemporary situations and opportunities were not included in the administrative registers but people with ADHD seemed to have a relatively higher risk (odds ratio: 1.55).

The study shows a low base rate of suicide attempts. Among 300,000 individuals born 1980 through 1985 followed from age 15 to their 29th births day

approximately 4.5 percent (13,358 persons) made a suicide attempt at least once during the period of 15 years.

About 57 percent of the population have a probability level less than 0.001 while 12 percent have a probability level higher than 0.003, and 6 percent of the population have a probability level higher than 0.004 (see Figure 1; Table 4). The results clearly indicate that our method to predict suicidal behaviour is much better than chance ($P<0.0001$) with cut-point between 0.001 and 0.18 (see Table 3). The problem is which cut-point should be used to give the most precise estimate.

The accuracy of the predictions follows a predictable pattern known from other studies in this field. It is nearly inevitable that the proportion of false positives will become high in this type of studies (Farrington, 2005). The criteria for high-risk level depend on the arbitrary choice of probability level (table 3). If a low level for high-risk is chosen for instance a probability level of 0.001, we will end up with a high proportion of false positives in the group of expected suicidal attempts (94.7 percent), but even if we choose a group with probability level of 0.01 (less than one percent of the population) we end up with 73.9 per cent false positives in the group of high-risk persons.

If a low probability level is chosen as a cut-point we will end up with a high proportion of false positive in the group of expected attempted suicides; if we choose a high probability level, we will have many false negative that is, many observed attempted suicides will not be recognized. In that case, treatment will not reach a fair proportion of the individuals who actually need treatment. In order to solve this dilemma, we use the ROC procedure (Receiver Operating Characteristic, ROC or Relative Improvement over Chance, RIOC).

The true positive rate (*sensitivity*) is defined as: a/(a+d) while *specificity* is defined as: b/(b+c) (Woodward, 1999). In table 4 we have chosen the probability cut-point 0.003, and on the basis of this compared *sensitivity* with *one minus specificity*. The *sensitivity* is clearly higher than one minus *specificity*. Since *sensitivity* is 0.6873 while *one minus specificity* is 0.2439 the prediction using the probability cut-point 0.003 is much better than chance ($P<0.0001$).

In Table 3 the sensitivity and the specificity has been calculated and the maximum of the sum-curve is at the cut-point 0.003 where the people most at risk of suicide attempt count approximately 12 percent of the population (see Table 4). Unfortunately, the number of observed attempts in the high-risk group only counted 12 percent while 88 percent turned out to be false positive.

One third (31 percent) of the actually observed attempted suicides will not be included in the treatment group if we use the most optimal cut-point with probability level 0.003. The relatively low accuracy of prediction follows a predictable pattern known from other studies in this field. It is nearly inevitable that the proportion of false positives will become high in this type of studies or we will end up with relatively many not treated in the group of expected (see Table 4).

We were not able to identify a minor group (less than 12 percent) in the population accounting for the majority of the suicide attempts. On the other hand, two thirds of the observed suicide attempts were included in the high-risk group.

Conclusion and discussion

An effective universal prevention strategy is changing or limiting access to a setting that may prompt, provoke, pressure, or permit an individual to engage in suicidal behaviour. This is corroborated by results in the present study where we find that adolescents with ADHD have an elevated risk of suicidal behaviour. Another universal prevention strategy is connected to certain characteristics of the local community. The assumption is that disadvantaged areas are having a low degree of integration. This is just speculative because the present study presents risk factors or precursors of suicide attempts, not causes. Results show that an indicator of moving into a disadvantaged area turns out to be a significant precursor of suicide attempts.

Many of the well-known risk factors are highly correlated with first-time suicide attempts. Young people attempting suicide experienced significantly greater amount of stress and chaotic disorder: Parental violence, long-term unemployment, parental mental illness, parental suicidal behaviour, disruptive families, and separation from caretakers. The results corroborate the assumption that disadvantages during the formative years are highly correlated with low self-esteem and hopelessness resulting in a suicide attempt.

The results indicate that the individual's decision-making processes in high-risk group explain differences in suicidal behaviour in the context of present and future resources and opportunities for education and income. Adolescents exposed to violence, anxiety, depression, severe stress reaction display an increased risk of suicidal behaviour.

On one hand, it does not seem possible to identify a minor group accounting exclusively for all of the observed suicide attempts. On the other hand, two thirds of the observed suicide attempts were included in the high-risk group (<12 percent of the population). These results give an increased knowledge of risk factors which can reinforce or improve efforts targeting at-risk groups. These results support the recommendations of the elements in the former Danish action plan (Nordentoft, 2007) and allow for the possible complementarity of population and targeted measures. A broader population strategy may be appropriate as a supplement to the targeted measures in high-risk groups, since one third of the observed suicide attempt derive from low- to moderate-risk groups.

Table 2. Estimated prognoses for the first time suicide attempt the following year. Discrete time Cox-model

Covariates	Type	% of Person-years	Model 1 (adjusted)			
			Estimate	SE Standard Error		Odds ratio
Constant term			-7.3342	0.0446		
16 year	0	7.4	-0.1476	0.0546		
17	0	7.4	0.1153	0.0506		
18	0	7.3	0.1778	0.0494		
19	0	7.3	0.2621	0.0481		
20	0	7.3	0.2402	0.0481		
21	0	7.3	0.1490	0.0488		
22	0	7.2	0.2135	0.0480		
23	0	7.2	0.1331	0.0486		
24	0	7.2	0.0767	0.0490		
25	0	7.2	0.1315	0.0484		
26	0	7.1	0.0820	0.0488		
27	0	7.1	0.0410	0.0491		
28 year	0	7.1	-0.0315	0.0498		
Contemporary situation and opportunities						
ADHD	3	2.2	0.4373	0.0311	***	1.549
Local community, social control, values and norms						
Disadvantaged area	2	1.6	0.2599	0.0526	***	1.297
Non-Danish	2	7.3	0.1227	0.0317	***	1.131
Developmental theories, disadvantages during the formative years						
Parental inpatient mental illness	1	12.2	0.2267	0.0240	***	1.255
Parental suicidal behaviour	1	4.2	0.1395	0.0345	***	1.150
Parental violence	3	12.0	0.1594	0.0229	***	1.173
Parental unemployment >21 weeks	3	64.7	0.1157	0.0213	***	1.123
Family separation	3	34.4	0.1879	0.0193	***	1.207
Child in care	3	3.1	0.2543	0.0302	***	1.290
Decision-making processes in high risk groups						
Not graduated from high school	1	55.6	0.5032	0.0228	***	1.654
No vocational training?		29.0	0.7453	0.0195	***	2.107
Victim of violence/sex offence	3	0.5	0.4384	0.0758	***	1.550
Anxiety	2	0.6	0.1805	0.0515	**	1.198
Depression	2	0.7	0.6653	0.0740	***	1.945
PTSD	2	0.8	0.6866	0.0734	***	1.987
Alcohol abuse	3	5.5	0.8833	0.0283	***	2.419
Drug abuse	3	2.4	0.5619	0.0327	***	1.754
Somatic disadvantages	1	20.7	0.3382	0.0189	***	1.402
Girl=1	1	48.8	0.4247	0.0179	***	1.529

Type 0: dummy variables included in the model. Type 1: risk factor observed at time t it also covers the years before and after the years under investigation. Type 2: risk factor observed at time t then the risk factors is also present at t+1. Type 3: risk factor observed at time t then risk factor is also present at all the following years. * $p<0.01$; ** $p<0.001$; *** $p<0.0001$;

Table 3

Classification table over probability cut-point estimated expected risk of first-time suicide attempt and observed first-time suicide attempt.

Cut-point Probability level	Thru positive: observed expected (a)	Not observed, not expected (b)	False positive: not observed, expected (c)	False negative: observed, not expected (d)	False positive[1] c/(a+c)%	False-negative[2] d/(b+d)%	a/(a+d)> c/(c+b) P<
0.000	13,358	0	286,088	0	95.5	-	-
0.001	13,064	52,215	233,873	294	94.7	0.6	0.0001
0.002	10,989	171,563	114,525	2,369	91.2	1.4	0.0001
0.003	9,181	216,309	69,779	4,177	88.4	1.9	0.0001
0.004	7,828	241,774	44,314	5,530	85.0	2.2	0.0001
0.005	6,736	255,202	30,886	6,622	82.1	2.5	0.0001
0.006	5,865	262,938	23,150	7,493	79.8	2.8	0.0001
0.007	5,206	268,139	17,949	8,152	77.5	3.0	0.0001
0.008	4,678	271,162	14,926	8,680	76.1	3.1	0.0001
0.009	4,285	273,109	12,979	9,073	75.2	3.2	0.0001
0.010	4,001	274,755	11,333	9,357	73.9	3.3	0.0001
0.012	3,564	276,913	9,175	9,794	72.0	3.4	0.0001
0.014	3,253	278,376	7,712	10,105	70.3	3.5	0.0001
0.016	3,030	279,569	6,519	10,328	68.3	3.6	0.0001
0.018	2,807	280,457	5,631	10,551	66.7	3.6	0.0001
0.02	2,615	281,265	4,823	10,743	64.8	3.7	0.0001
0.04	1,614	284,451	1,637	11,744	50.4	4.0	0.0001
0.06	1,123	285,227	861	12,235	43.4	4.1	0.0001
0.08	828	285,527	561	12,530	40.4	4.2	0.0001
0.10	600	285,715	373	12,758	38.3	4.3	0.0001
0.12	421	285,849	239	12,937	36.2	4.3	0.0001
0.14	297	285,943	145	13,061	32.8	4.4	0.0001
0.16	202	286,002	86	13,156	29.9	4.4	0.0001
0.18	129	286,040	48	13,229	27.1	4.4	0.0001

1) False-positive: the numbers of false-positive is calculated on the basis of the total number expected i. e. b/(a+b). 2) False-negative is calculated in per cent of the total number not expected, i. e. c/(c+d). 3) With cut-point between 0.001 to 0.18 predict suicidal behaviour better than chance (P<0.0001). Fishers exact test find that a/(a+d) > c/(c+b).

Note: Regression is based on a logistic 'discrete time Cox model' (Allison, 1982). Number of persons in the study n=299,446, followed from age 15 to the age of 29 years.

Table 4. Classification table over probability cut-point: estimated percentage of the population, and observed and expected first-time attempters (selected cut-points).

Cut-point	Percent of population	Observed, expected of expected a/(a+c)	Observed, expected of observed a/(a+d)
0.001	43	5.3	97.8
0.002	24	8.8	82.3
0.003	12	11.6	68.7
0.004	6	15.0	58.6
0.005	4	17.9	50.4
0.006	2	20.2	43.9
0.007	2	22.5	39.0
0.008	1	23.9	35.0
0.009	1	24.8	32.1
0.010	1	26.1	30.0

The Flynn Effect in Denmark,1939-90
Rise of the little guy

Thomas Lill Madsen

Introduction

The Flynn Effect is the increase in avg. population IQ by 20-25 points over the last 150 years in modern societies. In Danish Military data for the birth cohorts of 1939-1990 ('Børge Prien Prøven' ~ BPP), bpp test score[1] reached a ceiling for the cohorts born around 1980 after an increase of 7 points and a contraction of standard deviation from 13 to 10. The sigmoid time trends in the bpp and related variables indicate that the process has been going on further back in time and that we at the most are seeing the latter half of the FE.

There were two camps in explaining the FE: A. 'It is education' (Flynn) and B. 'It is nutrition' (Lynn) with years spent in school ('yos') and height as proxies for the mental and physical dimension. I aim to give each it's due in a model for *the Flynn Nexus* comprising the cohort variables: {bpp, stdev. of bpp, height, years of school ('yos')}. To these I add two further covariates, infant mortality and urbanism: Infant mortality as a proxy for physical health and urbanism via its relation to educational achievement.

'More education' is the factor with the most immediate appeal to the uninitiated as a potential cause for the FE, but this relation is bidirectional: intelligent people seek education and education has a reciprocal effect on IQ. For educations true effect on IQ we need data for an intial IQ, followed by an amount of education, and the 'resulting' IQ.

Whereas yos has the same meaning within and between cohorts, height is different. Within a cohort, height differences are to a large degree genetic in origin and of no interest in this connection. Between cohorts, height differences are entirely environmental and the gap between average height at time *t* and its fully grow value of ~181 cm is a proxy for cohort health or environmental deficits in a broader sense.

The current status of the FE has perhaps implications for the future of the low intelligence (IQ~85-90) of conscripts from the Middle East and Africa who, all though enjoying the good life in Denmark, have barely progressed IQ-wise from their countries of origin. Is there still a FE waiting to happen for them?

Data

A. Børge Prien Prøven data samples from cohorts 1939-1958 (N ~ 3300/cohort) and 1968, 1979,1984 from *Danish Army/Recruiment* were made available with thanks

[1] For *intelligence* I will be using 'bpp' when the reference is to the raw score from army data and 'IQ' as the more general term.

to Tom Teasdale. Fields: {bpp test score, height, years of school, year of birth,year tested}. For the period 1939-1958 *years of school* had an idiosyncratic coding which was converted to approximate number of years. For the cohorts of 1968-1983 *height* was by the most unfortunate assumption of its irrelevance deliberately deleted for these samples so average height had to be estimated by trend extrapolation.

B. For the test of 2009 (~ cohort of 1990) a full dataset from BPP-data (N~ 27000) including all variables in a research assignment for the Danish army. In conjunction with a database of surnames and their most probable country of origin (www.thomaslillmadsen.dk/ A surname model for ethnic origin) it was possible to differentiate between the ethnic groups, of which only 'Danes' was selected for compatibility with older data where the representation of immigrants was small.- Data for testscore, height, and yos at *the municipal level* were extracted and joined with GPA from a exogenous dataset.

C. For the historical period extending to 1862 only *height* was known from military registers for select years. For illustrative purposes, data for other variables were extrapolated by their relation to height in the 1939-90 period. This eg. resulted in the prediction yos = 7 for the cohort of 1862.

D. Macro data for urbanism and infant mortality were available for the long term prior to 1970 from S.Å.Hansen (1970) and DST (1990). 'Urbanism' defined as LB/(LL+LB) where LL: farm labor, LB: city labor. Extrapolated by logistic modeling to 1971- 1990.

The Flynn Effect in Danish Military Testing.

Presentation of data, fig.1-3

The causes for the FE must be of either mental or physical origin. These have in the literature been represented by education (Flynn's model) and height (Lynn's model) as proxies.- The present aim is twofold: 1) to establish underpinnings for these two factors and then 2) model bpp for the period 1939-90.

The FE could very well have been operating over more 130 years spanning an IQ difference of 20 pts. or more. We have records for the births cohorts of 1939-1990 from Army/personel ("Børge Prien Prøven") and can backcast these time series to earlier times in an attempt to visualize the full picture. All variables move in tandem over time in a semi-sigmoid fashion. *Height* is known for all years from other military archives and predictions for other variable are modelled from the linear relation to height observed in the later period. (fig 2.)

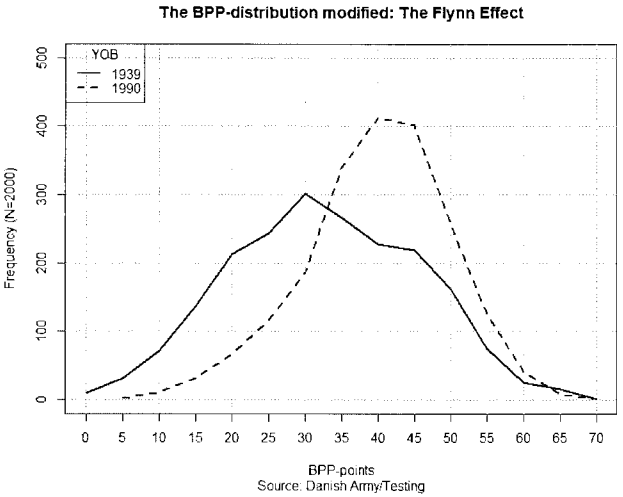


Fig 1. The Flynn Effect 1939-90. *The average scores were 35 (13) and 42 (9.7) respectively. At genius level, frequencies were constant.*

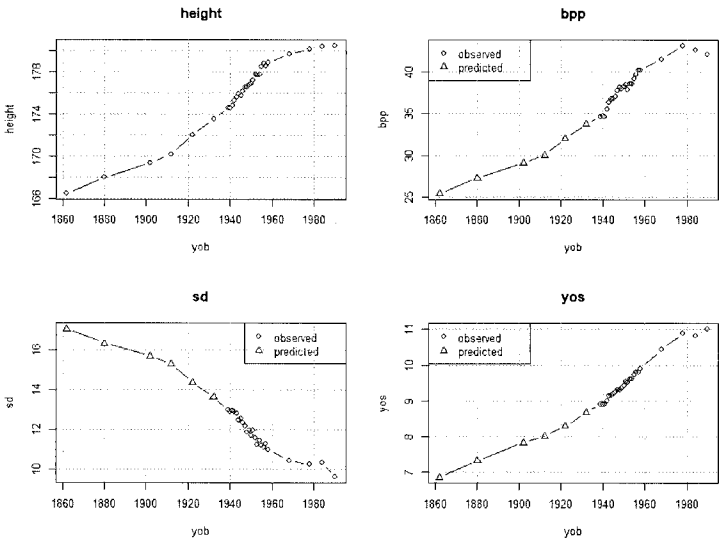


Fig. 2 The Flynn transition. yos: years of school, **sd**:standard deviation of bpp
Further covariates are *infant mortality* as proxy for physical health and *urbanism* related to education. Viewing the set of variables synoptically, the *Flynn Nexus* arises:

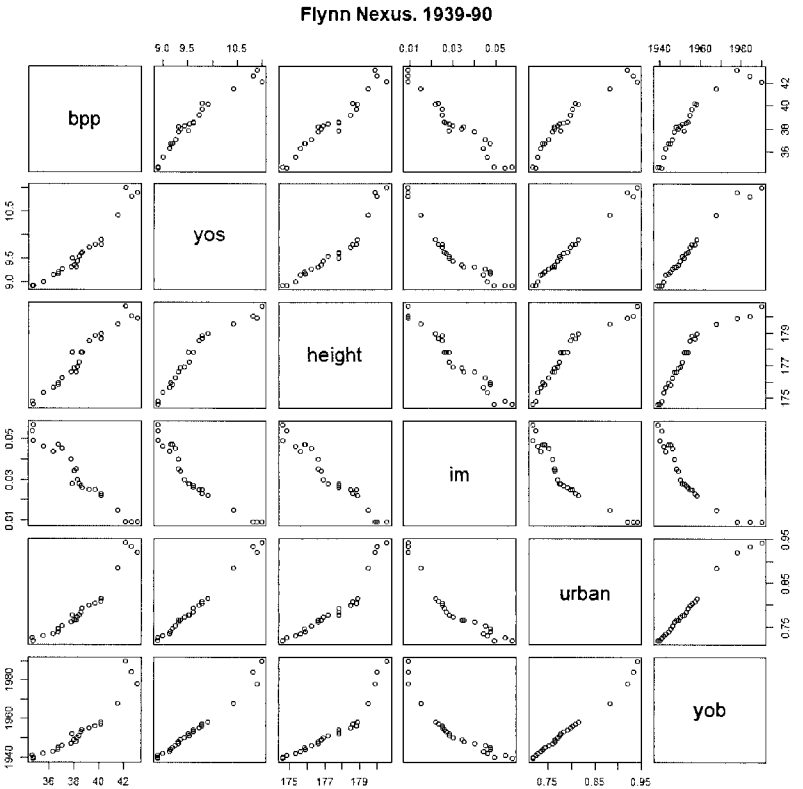


Fig. 3 The Flynn Nexus. For a given variable, read column-wise to see it as x-variable and row-wise to see it as y-variable. Eg. the 'Flynn- and Lynn models' appear in row 1, column 2 and 3. In RC(2,5): yos vs. urbanism. Time-trajectories of all variables in column 6.

2.1 Model for height 1862-1990.

Chort height has risen ~15 cm since 1862 in a semi sigmoid trajectory. Height and infant mortality ('im') as a proxy for 'health', are closely related with im exhibiting the mirror image of height. Using the logit-transform we estimate the model:

```
logit((height-166.2)/14.5)  = yob + 0.5/(0.5 + im):
                    Estimate Std. Error t value
(Intercept)            -3.04       0.28   -11.0
'yob'                   0.18       0.09     2.0
(0.5/(0.5 + im))        3.89       0.36    10.8
Rsq: .99   ('yob' = (yob-1862)/128)
```

Rsq must be taken with a grain of salt, but it appears that height in the period 1939-90

was 99% predictable from its relation to im estimated from only the previous 6 observations from 1862-1932, fig 4.3

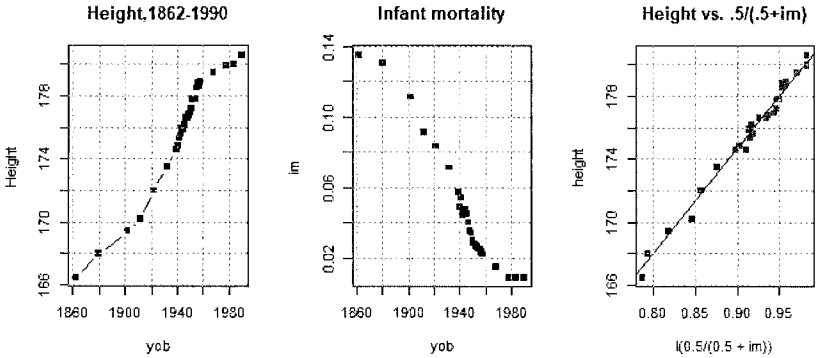


fig 4. Height of conscripts and infant mortality for the cohorts of 1862-1990

WW2 years of economic stagnation etc. in red- the upward trend was unaffected. fig. 4.3. Height 1939-1990 predicted (green) from the relation 'height = k/(k+im)' estimated on data from 1862-1932, with k = 0.5.

2.2 Model for years of school (fig.5)

A link between *yos* and *urban* could be mediated via a) absence of alternative occupation on the farm and b) easier access to education in cities.

	Estimate	Std. Error	t value
Intercept)	2.22	0.13	16.8
urban	9.35	0.17	56.0

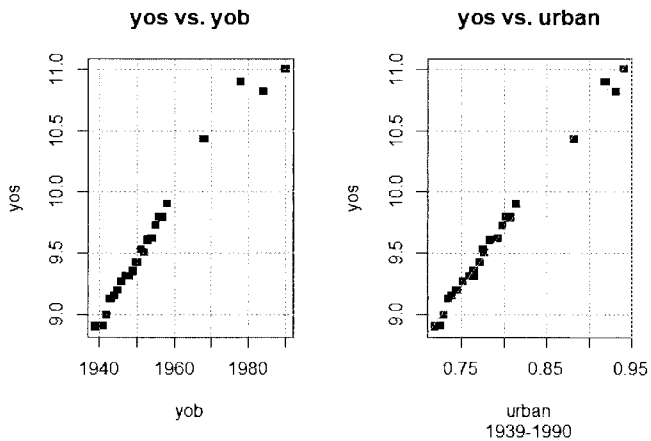


Fig. 5 Years of school (yos) well predicted by urbanization.

2.3 Flynn Effect model for average cohort bpp.

With data for height and education thus connected via the social transition from agrarian to modern, it seems justified to use both of these variables in a model for the FE. Other authors have naively estimated: 'bpp= f(yos)". But this is unsolvable within this dataset because of the interdependency of bpp and yos both at the individual and cohort level: a) Intelligence predisposes for education and b) education furthers intelligence. For b) estimates can be gained from either twin data or other data where IQ is measured before and after a given amount of education. -From twins with the bpp-dataset born in 1990, modelled as: bpp.twin2 = k1*bpp.twin1 + *k2*delta(yos), k2 emerges as 1.1 (.33), (~1.5 'IQ'-wise). We should therefore be prepared for an yos-coefficient of that magnitude. (cf. also: www.thomaslillmadsen.dk/the-iq-effect-of education). For yos, we are implicitly modelling against the alternative of "not going to school" which could have ment different things over the years. (Larger values for k2 were found by other writers).

For the period 1939-58 the linear relation between yos and bpp was exceptional regular (fig.6), but subsequent data deviated from this early trendline whereas the linear relation to height was quite stable over time. We can model bpp as: bpp = height + yos, with yearly data. (The data points are unequally spaced: for yob =1939-58 one year, for yob>1958 5-10 years, so they should probably be weighted proportional to 'delta years'.

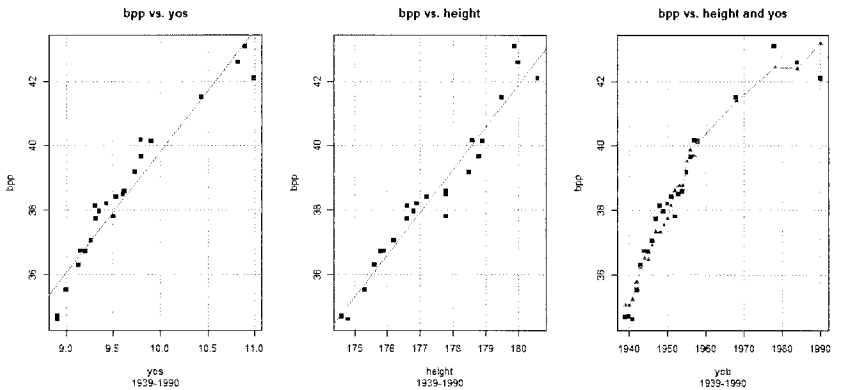


Fig 6. BPP-testscore vs. yos and height.

	Estimate	Std. Error	t value
(Intercept)	-144.84	27.40	-5.29
height	0.97	0.18	5.41
yos	1.09	0.49	2.25

R-sq: 0.98

Due to the high degree of multi-correlation between cohort yos and height, these

coefficients cannot immediately be taken at face value. The literature has estimates of schooling effect on IQ (ie.dIQ/dYOS) in the wide range of 1 to 4. (These would apply to normalized IQ, ie. stdev. =15. In bpp stdev is smaller,10 to 13, ie. the value of one bpp-point is higher) . [This model might, despite of irregular time intervals, have benefitted from a time series approach]

An alternative way of checking the yos-coefficient within the army data is to perform *regression by municipal districts* ('kommune', N=98). Bpp-data 1990 are then summarized by kommune and joined with the covariate of *GPA-testscore* ('karaktergennemsnit') available from another context for year 2004, end of 9th grade (compulsory), age ~16. It will generally not be the same students who took the Army test in 2009 but they were drawn from the same pool of genes.

At this level I suggest the model: bpp = yos + GPA, with municipal GPA as a control for prior intelligence or as accounting for municipal heterogeneity. To validate this factor, an estimation of 'bpp = b*height' on individual data (N = 70000) delivers b = .18 and we would like to find a similar coef. from the municipal data using GPA :

	Estimate	Std. Error	t value
(Intercept)	-47.07	19.03	-2.47
height	0.20	0.12	1.67
GPA	6.56	0.55	11.84

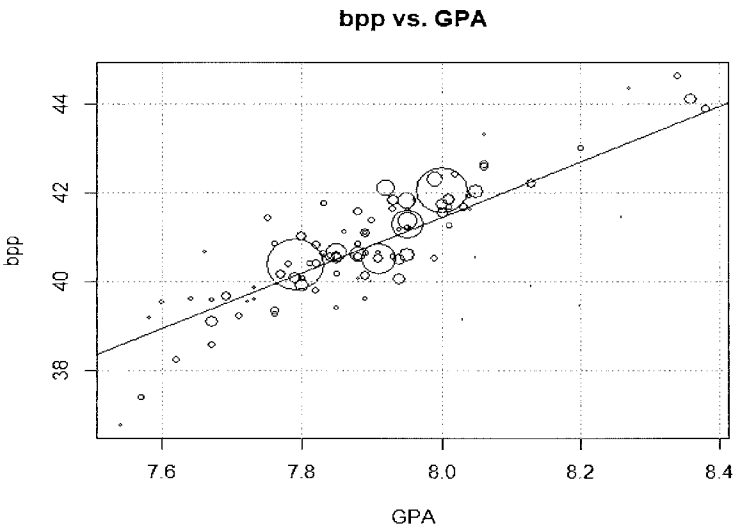


Fig. 6X. Municipal bpp predicted by GPA

This fails to reject GPA as a valid covariate. We can therefore proceed to an alternative estimate for the effect of education on bpp-score:

	Estimate	Std. Error	t value
(Intercept)	-14.606	2.985	-4.89
GPA	5.828	0.563	10.35
yos	*0.854*	0.258	3.31

Thus we find roughly the same coefficient (.85 vs. 1.09) for yos in the two models. Note the high predictive power of GPA.

3. Rise of the little guy

It is universally recognized that height and IQ are positively correlated but how strongly and what is the ontological basis for the relation? Are there a set of genes coding for both brain size and height? Or micro environments impacting both variables in the same direction?

At the individual level, the relation between height and bpp-score has changed considerably over the period. In the 1939 cohort, height beta, ie. the bpp-equivalent of 1cm of height, was 0.55 but diminished to .18 in 1990:

	Estimate	Std. Error	t value	
(Intercept)	-61.79525	6.76525	-9.13	
height beta	**0.55463**	0.03873	14.32	
I(yob = 1990)	1.36776	0.17942	7.62	'time effect'
height:I(yob - 1939)	-0.00718	0.00101	-7.08	~ -.37

This represents a huge change reflected in fig.7, which graphs bpp-scores per height interval in the years 1939 and 1990. As the average height converged to ~180.5cm, height beta was reduced from .55bpp/cm to .18bpp/cm.- Why was this? Yos increased relatively for the little guy but insufficient to account for more than (1/3) of the change.- If cohort H has reaching 181cm entails that *environmental insults* have seized to be an agent of differentiation, only genetic or random differences remain. One would think that the little guy- to the extent that height overlaps with social status- would have been hit harder thru time, and if so, standard deviation of height should have decreased but this has not happend, cf. fig. 8, which displays

samples of 2000 probands for each year sorted by respectively height and bpp-score.

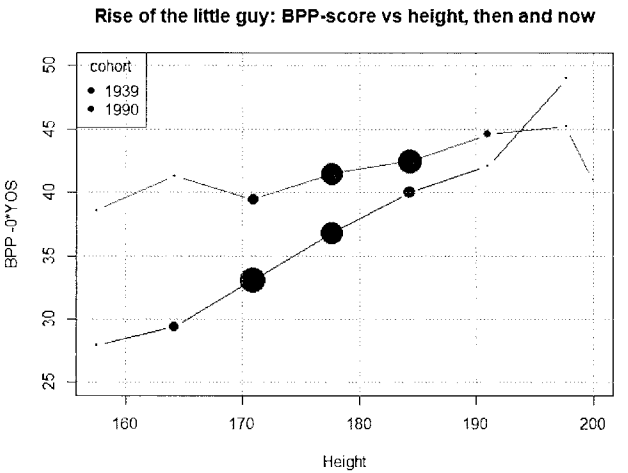


Fig.7 Height and BPP-score (a 5cm height lift is visible)

Within twin couples born in 1990 (N=259 pairs) height differences had undetectable effect on bpp.- If cohort H reaching 181cm entails that *environmental insults* have seized to be an agent of differentiation, only genetic or random differences remain. **fig. 8**, displays samples of 2000 probands for each year sorted by respectively height and bpp-score (median value for each year can be read off the y-axis for rank=50%). For height, the curves have risen in parallel for more than a century and presents in comparison with bpp-percentiles a different picture. Surprising that the environmental deficits that lead to subpar height overall had a democratic distribution initially: the top has gained as much as the bottom with standard deviation constant at approx. 6.6 cm. such that the total increments amounts to 2.4 standard deviations.- A person of average height born in 1990 would look like a giant amongst dwarfs in 1874. In contrast, for bpp-score the curves converge left to right, implying that the bottom has risen relatively but *at the high end, all years are the same.* Thus stdev of bpp has declined.

4. The ethnic difference

Probands born in 1990 and of Middle Eastern, African and Tyrkish background have an avg. height of 175 cm and bpp-score of 33 (10). For Nordic origin bpp score is 42 (10) [2]. The future of this IQ-gap is of some interest for the long term as the population with ME etc. background in Denmark is doubling every (~50) years by

[2] *Asian* probands were roughly similar to Danish.

immigration and differential fertility. Is there an *intrinsic* ethnic difference or just a 50-75 year time lag.

A difference in both IQ and height could arise on several accounts: a. consanquinity ('inbreeding') b. dietary habits or c. genetic ('intrinsic', 'essential'). In the case of a. and b. differences should have a tendency to diminish. Consanquinity can only account for at very most 2 bpp-pts.- Predictions on the basis of the equality of the standard deviations indicate that scores should already be level, ie.10 by which the FE seems to have run its course. This is seen clearly that the rank-plot in **fig. 9** is substantially different from the bringing-up-the-rear situation in fig. 8. First of all, there is an almost complete lack of superior intelligence (bpp > 59) in the ME-group. Further, the gap between the curves hardly narrows, left to right- in contrast to fig.8. This could imply a more permanent difference in line with Lynn (2002) for ME-countries.

Conclusion

In Denmark, the operating principle of the FE was to increase the mean IQ towards an asymptote *and* at the same time reduce the standard deviation by a rightwards compressing on the distribution. The net result being that the mean score has increased by lifting the bottom while keeping the very top stable: genius frequency is unaltered throughout but low intelligence is made scarcer. In a traditional 'IQ'-measurements the increase in the mean would be amplified by the simultaneous contraction of the standard deviation.

Out of the total bpp-rise of 7 points between 1939 and 1990, two extra years of school contributed 2.2 points, and the remaining 4.8 points were attributable to increases in height. This measure correlated higher than any other variable with IQ over time and can, as an empirical proposition, account for the remaining 2/3, even though the details of the causal link will remain unclear. Whatever caused height to increase- 'health care and nutrition'- helped the lower ranks disproportionally IQ-wise and as environmental deficits disappeared as source of variation, standard deviations of IQ receded towards a level defined, perhaps for the first time in many centuries, by genetic heterogeneity.

Other estimates for The Flynn Effect over the 20th century fall in the range of 2-3 IQ-points per decade. Compared to this and for the period 1890-1990, cf. fig. 2, I estimate a difference of (42-29) = 13 bpp-points over 10 decades. Converted to IQ-score from the latest bpp standard deviation of 10 for yob = 1990, this equals 2 IQ points per decade (~13*15/100), ie. within range.

For Flynn, reasoning within in a one-parameter model to account for a two-parameter event, it appeared paradoxical that backcasting IQ from current trends resulted in such an incredulously low IQ three generations ago.- We can now see how this apparent paradox arises when scores are rising and standard deviations are

declining. If the new generation has testscore ~ ND(42,**S**) and the class of 1890 had test score ~ ND(29, 16), then the smaller the value for **S**, the dimmer the old generation appears, since *we* judge it by (42-29)/**S.** For S = 10, this difference would amount to 15* (42-29)/10= 20 IQ-points. *They*, in turn, would bow to us, but only in the order of 15*(42-29)/16, ie. 12 IQ-points,- not quite so *massive*.

But who would think of re-norming the *cm-scale* because conscripts are getting taller? Likewise, the standardization procedure routinely applied to IQ-testing is disinformative. The proband who in 2009 scored 42 on the BPP-test is intelligence wise at par with his fathers grand daddy if he in 1909 had scored 42,- whatever the standard deviations were at that time.

Perhaps the cultural renaissance that Flynn was looking for, never happened because the creative top segment has been stable in proportion. Flynn also encapsulated his effect by the saying "All boats rise on the tide". An evocative image, but by far an apt description of Danish data. Can we really believe that the mechanism of the FE is fundamentally different between similar countries? That while the scenario in Denmark and Norway is 'bringing up the rear', but in Sweden it is 'all boats rise'? Since the countries in Western Hemisphere, and in Scandinavia in particular, have developed very much in parallel, it seems unlikely that they should be different in their essence in this respect. But in the British tradition of empiricism, even grains of sand is special and unique.

In conclusion, the Flynn Effect seemed to 1/3 about more education and 2/3 about neutralizing environmental deficits. It is befitting that such a simple account can be given now that it is all over.

References

1. William, Robert L. (2013) *Overview of the Flynn effect,* Intelligence 41 (2013), p. 753. (with bibliography)
2. Lynn, R. (2009) What has caused the Flynn Effect? Secular increases in the development quotients for infants. *Intelligence 37 (2009), 16-24*
3. Flynn, J.(2009). Requiem for nutrition as the cause of IQ gain: Ravens's gains in Britain 1938-2008. *Economics and Human Biology (2009) 18-27*
4. Jensen. A.(1998). The g-factor
5. Woodley, M.A (2009): Inbreeding depression and IQ in a study of 72 countries, Intelligence 2009, p. 268-276.
6. Hansen, S.AA.(1976), *Økonomisk vækst i Danmark* ["Economic Growth in Denmark"]
7. Steel,D.(yyyy) Logistic regression analysis with aggregate data: tackling the ecological fallacy.
8. BPP-data for various years, thanks to Danish Army Recruitment and Tom Teasdale

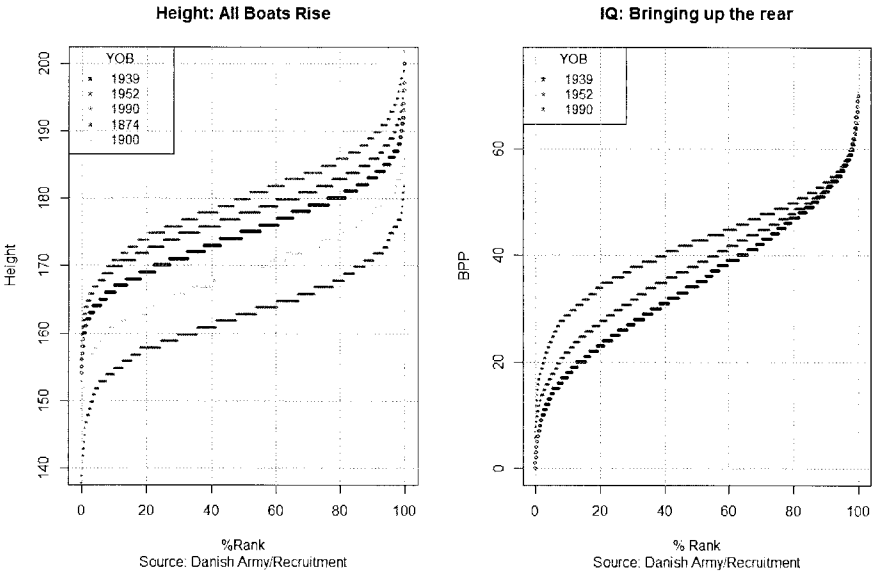


Fig.8 Height and bpp-score percentiles (in 8/1: disregard signature).M edian value for each year can be read off the y-axis for rank=50%).

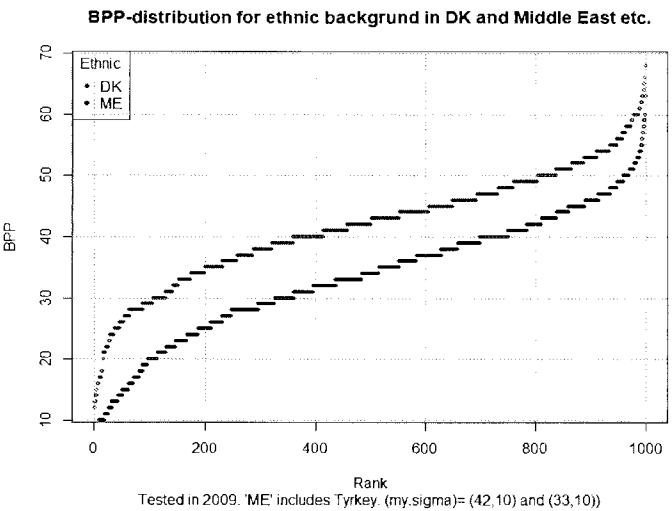


Fig.9 A Flynn Effect apart? Bpp-distribution for ethnic Danes and immigrants with the Middle Eastern and African background.

Intervention Models for Sales of Cars in Denmark

Anders Milhøj

Department of Economics

University of Copenhagen

anders.milhoj@econ.ku.dk

Abstract

Cars in Denmark are heavily taxed, and the level of taxation has changed from time to time. In recent years first an extra tax was introduced in order to reduce fuel usage and two years after a new tax was introduced in order to make more secure cars affordable. The first interaction was in favor of small cares while the last intervention was in favor of large cars. It differs whether such changes in the taxation was introduced immediately or were announced a few months before. Also changes in taxations are expected by everyone interested in the negotiations in parliament.

Statistics Denmark publishes the number of sold cars in many segments. In this paper the car segments small, medium and large are used in order to illustrate the effect of these changes in taxation.

The models applied are estimated by PROC ARIMA with intervention components, as exogenous, deterministic effects like changes in the taxation should be modeled separately by deterministic components and not as a part of the stochastic model. In the present paper models for exponentially decreasing impacts of an intervention are of interest. However, also situations were impacts are increasing up to the actual date of an announced future change in the taxation exist in this data set. Such situations are easily modeled by reverting the direction of time.

Introduction

All time series models have a stochastic component to fit the influence from various sources outside the model. These impacts are in a statistical model seen as outcomes of stochastic variables. They are often denoted "noise" which is an intuitive explanation for the fact that they are seen as noise that makes the true signal, the correct data series, more dim. Often the series is however also influenced by events with large impacts on the time series. Such events are explicitly known as extraordinary.

We reserve the concept "interventions" to reactions to events that are of this non-stochastic nature. The existence of an exogenous intervention is known to the analyst and its impact should be modeled separately from the stochastic model of the "remainig" data series. The reaction to the time series of the intervention has some kind of deterministic explanation so the behavior of the series close to the time of the intervention has to be modeled apart from the stochastic mechanism that otherwise generate the observations. Otherwise the estimated residual variance become unrealistic large.

An example of such an event is the intake of medicine where biological measurements indicate a reaction to the intake. It is known that the medicine is given so any change in the physiological behavior of the patient is assumed to be due to the medicine. Another example is the reaction of the consumers to an increase in taxes, where the consumption is believed to decrease.

Intervention models are often formulated by dummy variables taken values like zero and one and hence they are by no means observed values. An intervention could model the level shift of a time series from summer to winter, but actual measurements of temperatures, humidity, hours of sunshine are by no way included in the model. For the two intuitive examples mentioned above this means that the actual dose of the intake and the actual level of the tax increase are not included in the model. More refined models for interventions are however very close to models relying on actual observed values of the dose or taxation percentages so the distinction is by no way strict.

A general framework for analyzing time series were presented in the book by Box and Jenkins(1976) and after that seasonal ARIMA models are standard when modelling univariate time series. Box and Jenkins(1976) in their book also considered intervention models of the kind considered in this paper.

Types of Interventions

Interventions may take various forms, as for instance one particular outlier, a level shift or perhaps a shift in the observed trend in the time series. More complicated interventions exist for instance a shift in the level which is not abrupt but instead occurs gradually.

An intervention for an intervention that only affect the series for just one observation, at time t_0, has the form

$$X_t = \omega I_t + Y_t \tag{1}$$

where the intervention components I_t is given by $I_t = 0$ for all t but for just the value $I_{t0} = 1$ at the event of the event. More precisely $I_t = 0$ for $t \neq t_0$ and $I_t = 1$ for $t = t_0$. The remainder term Y_t could be modeled in many ways according to the situation, e.g. as a

regression model or as a time series model such as a multiplicative, seasonal ARIMA model, see Box and Jenkins(1976).

For interventions for temporary level shifts we assumed above that the effect of the intervention was immediately, but often the effect lasts for some time. This could be modeled by introducing lags in the intervention model

(2) $$X_t = \omega_0 I_t - \omega_1 I_{t-1} - .. - \omega_r I_{t-r} + Y_t$$

Here the effect of the intervention is given as steps in the following way

Time t_0 : The effect is ω_0

Time t_0+1 : The effect is $-\omega_1$

..

Time t_0+r : The effect is $-\omega_r$

The choice of sign for the parameters ω_i in this parameterization is taken as the notation by Box & Jenkins(1976).

One problem by using $r > 0$ in this parameterization is that the number of ω–parameters could be large and as they tell the same story they are subject to multicollinearity. This means that the estimates of the ω_i parameters have large standard deviations.

If $r > 1$ a more *parsimonious* parameterization could be often more efficient. In order to reduce the number of parameters it is often better to adopt another form of parameterization using the notation of lag polynomials. The idea is that the effect is slowly decaying by a factor δ_1 as

(3) $$X_t = \omega I_t + \omega\delta I_{t-1} + .. + \omega\delta^r I_{t-r} + .. + Y_t = \frac{\omega}{1-\delta L} I_t + Yt$$

Here the effect of the intervention is given as steps in the following way

Time t_0 : The effect is ω

Time t_0+1 : The effect is $\omega\delta$

..

Time t_0+r : The effect is $\omega\delta^r$

... etc.

The value $\delta = 0$ gives the simple intervention (1). But the situation of a positive parameter $0 < \delta < 1$ gives a gradual decrease towards the previous level of the series which is never reached but in practice the first steps are the largest. An example is presented in Figure 1 for $\delta = 0.8$. For $|\delta| \geq 1$ the series diverges and a new level is undefined. The parameter δ is therefore restricted to the interval $]-1, 1[$. Even if the value $\delta = 1$ is not included in the parameter space values of δ close to +1 takes the form

of a step function. These kinds of models are discussed in detail by Box and Jenkins(1976).

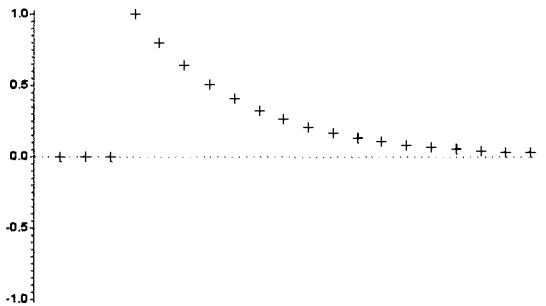


Figure 1. A Gradual Intervention Component for $\delta = 0.8$

Sales of new cars in Demark

Statistics Denmark publishes monthly data for the sales of new cars in form of many time series. In this paper we apply series for the sales numbers in three segments of cars for private use: Mini (like Volkswagen Up), medium (like Volkswagen Golf) and Large (like Volkswagen Passat). The time series are published with start January 2004 and data up to December 2017 is used in this paper - total of 168 observations.

The Danish taxation on new cars is heavy. For cheap cars a taxation is at present 85% is added to value, while the taxation of the values higher than a certain limit, around 185.000 DKK, is as high as 150%. Previously these percentages were even higher 105% resp. 180% and the limit was lower than now, around 85.000 DKK. On top of all specific taxes the general value added tax adds 25% to the final price of all commodities.

By definition this taxation system favors sales of smaller, cheaper cars. But in recent years the system has changed somewhat a new tax in 2013 depending on the fuel economy was been introduced, and in 2017 a new taxation (in fact a tax reduction) depending on security factors was introduced. While taxes depending on fuel economy favors smaller cars as they use less fuel, taxes depending on security favors large cars as it is safer to be a passenger in a large car in case of a frontal crash.

The sales series for the three segments series are plotted in Figure 2. Figure 3 shows average prices as met by the consumers in Danish Kroner for the three segments. The vertical lines on both plots mark some of the changes in taxation as an indicator of large interactions in the sales.

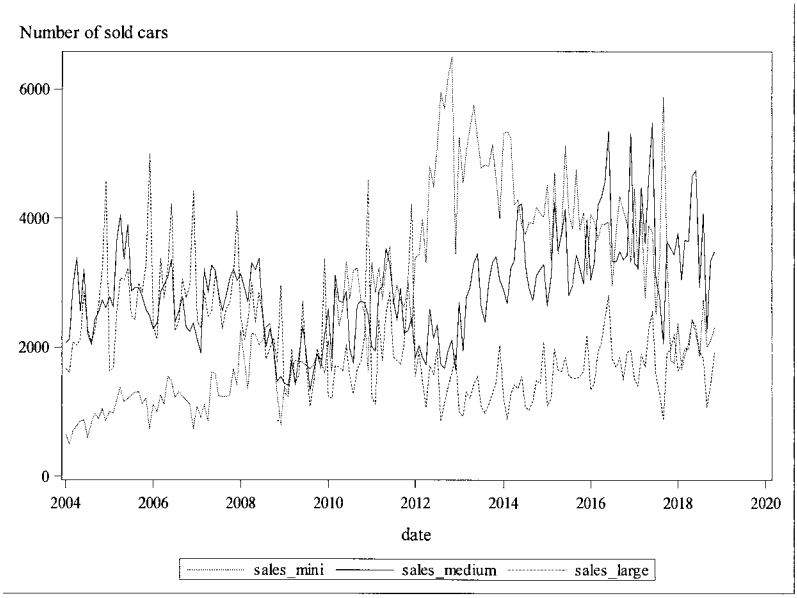


Figure 2. Monthly sales of cars in three segments

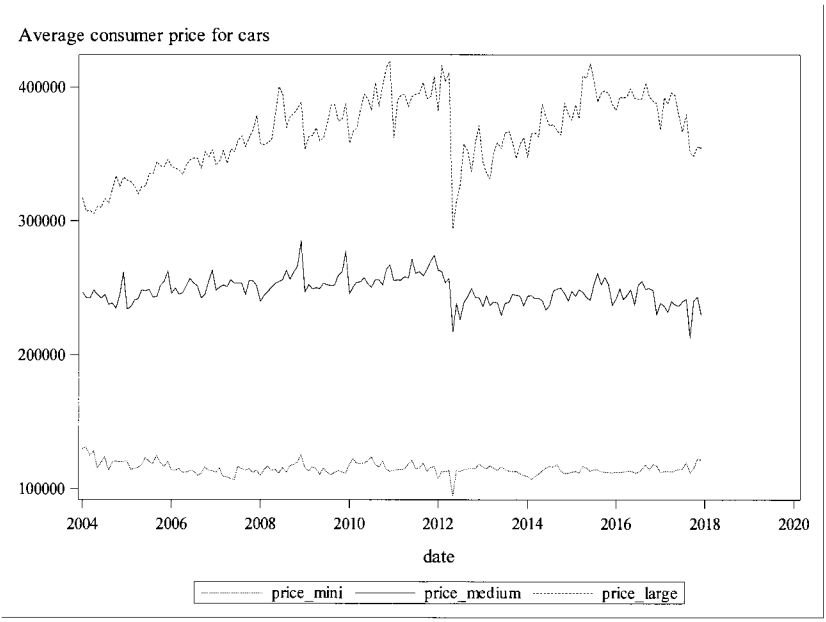


Figure 3. Average price of monthly sales of cars in three segments

ARIMA models for monthly sales of cars

As a benchmark model the so called airline model is used. This model is often applied as a default model for seasonal time series; made famous by the example of the number of airline passengers by Box and Jenkins (1976). The model is also applied as default in the X11-ARIMA algorithm for seasonal adjustments. For these datasets the model with parameters estimated by maximum likelihood are

Mini	$(1 - B)(1 - B^{12})X_t = (1 - 0.52B)(1 - 0.877B^{12})\varepsilon_t$, $var(\varepsilon_t) = 533.63^2$
Medium	$(1 - B)(1 - B^{12})X_t = (1 - 0.58B)(1 - 0.68B^{12})\varepsilon_t$, $var(\varepsilon_t) = 455.39^2$
Large	$(1 - B)(1 - B^{12})X_t = (1 - 0.66B)(1 - 0.61B^{12})\varepsilon_t$, $var(\varepsilon_t) = 397.58^2$

Automatic outlier detection

The variances are heavily influenced by extreme observations as they are estimated using the sum of squared residuals. Two times the standard deviation is of course a much too narrow limit for outlier detection as the expected number of outliers in that case is $0.05 \times 180 = 19$. The critical value for a 5% test for outliers among 168 observations is 3.92 times the standard deviation. However, the residual variance is rather large due to the existence of outliers, which are nor corrected for any interventions. So a limit of three time the residual standard deviation in the ARIMA models are applied.

For the three segments we find a few months where the residuals have absolute values larger than three times their standard deviation and only one is larger than four times its standard deviation. That is the sales in the segment mini for November 2017 which was 2532 higher than expected.

If these residuals are removed using dummy variables, the residual variance is reduced significantly. Of course more outliers could be compensated by dummy variables in order to reduce the variance even further, but the number of parameters then of course increases as each outlier leads to an extra parameter.

Interventions followed by an exponential decay

A numerically large residual could be considered as the simplest form of an intervention which is easy to see as a part of an ordinary time series analysis. Interventions followed by an exponential decay (3) are more advanced and not so simple to identify. But by macro programming it is possible to fit an intervention of the form (3) for each month in the dataset. However, the denominator variable δ is not well defined if the intervention is applied for the very last part of the series. The next Table gives the estimated values of ω and δ for the months were ω numerically exceeds four times its standard deviation and the t-value (estimated value divided by

its standard deviation) exceed three. In some situations, the estimated δ is close to one, which means the intervention is probably a step function. Several of the entries in table show values of δ not significantly different from 0.7, which are interventions of the form (3).

Sales of Mini					
Date	ω	δ	std dev δ	t-value δ	probt δ
01/07/2012	1673.81305	0.96556233	0.04398449	21.95	<.0001
01/12/2012	-1891.4035	0.61350024	0.18632232	3.29	0.0010
01/11/2017	-1949.9256	0.99917803	0.03640831	27.44	<.0001
Sales of Medium None					
Sales of large None					

Outlier detection in the reverted series

This model is fitted in order to take care of the autocorrelation structure in the time series. The estimated ARMA parameters and the residual variance has the same values when they are estimated based on the reverted time series. However, the idea of an innovation process, ε_t, and hence also the residual values are different. The residuals are now prediction errors when backcasting one period using only future values of the time series. In this case we also see outliers, but they are not necessarily for the same time stamps as outliers in the usual forecasts.

For the three segments we find a few months where the residuals have absolute values larger than three times their standard deviation and only two are larger than four times its standard deviation. That is the sales in the segment mini for September 2017 which was 3269 higher than expected and the sales for the segment Large for December 2011 where the sales were 1961 larger than expected.

Exponential effects before an intervention

If the reverted series are the in the model for exponential decay, we find interventions which is in actual time direction exponentially build up in the months before rather than decaying after the interventions. The following table gives the estimated values of ω and δ for the months were ω numerically exceeds four times its standard deviation.

For some of these interventions the estimated δ is close to zero. This means that the intervention this particular month is probably an outlier. For one month, the estimated

δ is close to one. This means the intervention is probably a step function. Several of the remaining entries in Table show values of δ not significantly different from say 0.7; that is of the form (3).

These results are based on an estimation of the denominator parameter δ for each month in the dataset. In order to ease the estimation burden a fixed value of δ, say δ = 0.7, could be applied so that only the ω parameter has to be estimated.

Sales of Mini

Date	ω	δ	std dev δ	tvalue δ	probt δ
01/12/2012	2354.52301	0.79254481	0.09190283	8.62	<.0001

Sales of Medium None

Sales of Large None

Statistics cannot replace knowledge!

In the analyses presented for this series up to this point interventions are detected automatically by a screening process. In order to see whether these findings are due to well defined events like e.g. changes in the taxation, we have to look through all history books and newspapers in the past many years for information of the history of Danish car taxes - and what else could affect car sales. Hmmm - sounds cumbersome, even though Google is your friend.

It is optimal if a statistical model includes all known external factors for car sales. A more honest analysis takes the approach first to identify all possible reasons for interventions with an ideas of the form of the intervention and after that estimate the parameters of the intervention components. Again - sounds cumbersome, even though Google is your friend.

If the purpose of the analysis is mainly focused on the most recent part of the time series a screening, like the one presented above is, however a practical approach to help the analysts to remember long forgotten events.

References

Box, G.E.P. and Jenkins, G.M. (1976), *Time Series Analysis: Forecasting and Control,* Revised Edition, San Francisco: Holden–Day.

Baby Patterns

A Mother's Guide to Understanding Her Baby

Sara Armandi

1 Introduction

Parenting life is tough. According to the National Diaper Bank Network, most babies are in diapers for 2-3 years. During this period a baby goes through 6-10 diapers a day. Using 8 diapers in a calculation, this corresponds to 2,920 diapers each year [NDBN, 2018]. Not only may this huge amount of diapers corrode wallets, but what is really disturbing, is the amount of time spent by parents on changing diapers. Unfortunately, not only diaper changes take up valuable time for parents. Other activities worth mentioning are time spent on trying to get an overtired baby to sleep, as well as the unimaginable additional time it takes to get out the door with all sorts of baby equipment.

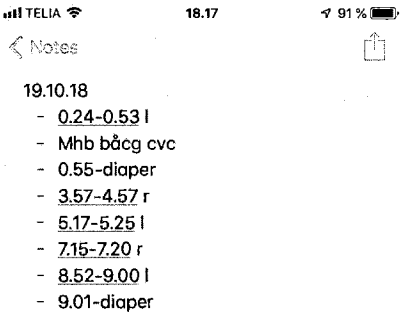
It can come as no surprise that parents occasionally may value baby free time. However, this time can be difficult to find, especially for mothers. On top of all the above-mentioned time-consuming activities, most babies in Denmark are exclusively breastfed the first six months of their lives. This puts extra strain on mothers, who babies are fully dependent on. Therefore, you will often find mothers who do everything possible to try to find some kind of pattern in their baby's behavior and needs. If there is just one time during the day, where mothers can get themselves a well-deserved break, and maybe even do something else than take care of their little nugget, they might avoid turning completely baby crazy. Parenting life is tough, especially motherhood.

In an attempt to help mothers, figure out their babies, this article seeks to find patterns as well as deviations within a baby's eating behavior. This is, among other things, done by using procedures from the SAS/QC® software, graphical displays and descriptive statistics.

2 Data

The analysis is based on data from one cute baby, collected over a four months period. Each time the baby started eating, the mother registered date and time, and likewise, when the baby finished eating. To simplify the data collection process, the registration of feeding date and time was done, using the Notes App on a mobile phone, see Figure 1. This raw data was then processed and imported to SAS.

Figure 1: Raw Data



Source: Screenshot from Notes App on mother's mobile phone.

The first observation was registered on August 30th, 2018 at 2:30 am, when the baby was three weeks old. Collection of data is still going. To uniformize the data, only observations from September 1st, 2018 to December 31st, 2018 are included in the analysis. In total, there have been 1,781 baby meals during this period.

2.1 Cleaning and Reduction

In Figure 1, an example of the unprocessed raw data is depicted. As can be seen, data does not only contain date and time each time the baby eats. In addition, it includes information on diaper changes, and from November 15th, registration of the baby's sleep is added. This information is initially deleted from the data. Further, Figure 1 clearly shows, that the data is not 100 pct. trustworthy. As data is manually registered at all times of the day, by a very tired mother, small peculiarities occur. Hence, data suffer from a mild degree of human error. However, this problem is immediately not a very frequent problem, why observations as "*Mhb båcg cvc*" in Figure 1 is disregarded, and the observations removed.

Finally, some feedings of the baby have occurred with only a few minutes apart and could therefore be considered as a single feeding. To account for this, all feedings with less than 30 minutes to the subsequent, have been deleted. This results in a total of 1,229 unique feedings, all at least 30 minutes apart.

2.2 Non-Equidistant Timestamp Data

In its immediate form, the data is non-equidistant timestamp data. It can also be considered as an unevenly spaced time series, which means that the spacing of observations is not

constant. This article attempts to analyze data in its unaltered form, and thus, not transform data into equally spaced observations, which is otherwise a common approach.

Three ways, to avoid the equally spacing transformation, is used. The first approach is to construct a new data set with a variable containing all datetimes from September 1st, 2018 at 00:00:00 to December 31st, 2018 at 23:59:59 with 15 minutes intervals. Next, a dummy variable, based on the 1,229 unique feedings is created. Each original time, for when the baby starts eating, is rounded to nearest 15 minutes. This is done using the SAS ROUND function in an unusual way:

```
Round(start_datetime,hms(0,15,00))
```

This results in a data set with two variables, the datetime variable with 11,712 observations, and a dummy variable, indicating whether the baby is eating or not at a given 15 minutes interval.

The second approach, is to construct a new variable which shows the time between two consecutive feedings. In SAS, this time between events can be calculated using the LAG function:

```
time_between = start_datetime - lag(end_datetime);
```

Finally, the third approach on which to handle the data, without transforming it to an equally spaced time series, is to aggregate times within the same hour. For instance, all observations that lies between 07:00 and 08:00 are assigned the same time unit. In this way, the number of observations for each hour is obtained.

3 Theory

A challenge with data as it exists, is the structure. Standard time series analyses are not applicable when using non-equidistant time data. Usually, data with unevenly-spaced time observations would be transformed into equally spaced observations.

One way to approach the unevenly spaced time series data, is by using the SAS/QC® software. Here, "*QC*" stands for *Quality Control*, and the software is often used to decide whether a process is in a state of statistical control [SAS/QC, 2018]. Below, the theory of two procedures available in SAS/QC®, which may work on analyzing the data, will be reviewed. In addition, a procedure from SAS/GRAPH® is also very briefly introduced.

3.1 Control Chart

Control charts, also known as Shewhart charts (after Walter A. Shewhart) are used to determine the variation in a process over time. A control chart consists of three lines, an average, an upper control limit (UCL) and a lower control limit (LCL). When a value is outside the control limits, it is said to signal unusual variation [SAS/QC, 2018, ch. 19].

One commonly used control chart is the *c* chart, which can be used to monitor counts of an event, often the number of defects. In SAS, this is created using the SHEWHART procedure:

```
proc shewhart data=feedings15;
     cchart feed * datetime / totpanels=1;
run;
```

In this procedure, `feed` is the process that should be analyzed, and `datetime` is the subgroup-variable. The `totpanels=1` option specifies that only one panel should be used to display the chart. The values of the subgroup-variable indicate how the observations in the input data set are arranged into rational subgroups. Examples of a subgroup-variable are dates or times at which subgroup samples were collected. In the data used for this article, the subgroup is a long sequence of time stamps, starting September 1st, 2018 at 00:00:00 with 15-minutes intervals until December 31st, 2018 at 23:59:59. In some of these 15-minutes intervals, the baby is eating. Each time the baby is eating, this is considered a nonconformity, compared to all other times. In this case, `feed` is a dummy variable, indicating if the baby is eating or not. On a *c* chart, each point represents the number of nonconformities for a particular subgroup [SAS/QC, 2018, ch. 19].

3.2 Rare Events Chart

The rare events chart is a specific control chart for rare events, that occur infrequently and with a low probability. What applies to the commonly used control chart are also applies for the rare events chart. The difference is the data plotted, as it in a rare events chart represent the times between successive events [SAS/QC, 2018, ch. 17]. In this example, when a baby is eating, is considered a rare event, even though it might not always feel like an infrequently event.

When creating a rare events chart, it is assumed that the events are independent. Hence, the occurrence of one event must not affect the probability of another, and the probability of an occurrence should be constant over time. This assumption is violated in the case of feedings, as a baby is no longer hungry when he has just finished eating, but he is very

hungry, if he has not eaten for a while. However, this is disregarded in this article, as the article is more considered as an example of how to use the procedure.

The rare events chart has an upper probability limit (UPL) and a lower probability limit (LPL). A value greater than the UPL indicates that the time between events is increasing, in which case the events are occurring less frequently. Conversely, a value less than the LPL indicates that events are occurring more frequently [SAS/QC, 2018, ch. 17]. The chart is created in SAS using the RAREEVENTS procedure:

```
proc rareevents data=feedings;
     compare time_between / dist=Weibull;
     chart time_between / dist=Weibull totpanels=1
          readphases=all outlimits=feedlimits;
run;
```

The `compare` statement produces a graph that compares the process data to a reference probability distribution. By default, the geometric distribution is used for integer data and the exponential distribution is used for continuous data. By specifying the option `dist=Weibull`, the Weibull distribution is used instead. The probability limits on the Weibull distribution is given in Equation A2 in Appendix. In this code, `time_between` is the variable containing time between events values (in this case the time between two feedings). Like in the *c* chart, the `chart` statement produces the chart, using the variable, `time_between`, which is the process-variable containing measures of time between events. The option `readphases=all` selects predefined phases from the data set for processing, whereas `outlimits=feedlimits`, creates a data set that contains probability limits for the chart [SAS/QC, 2018, ch. 17].

An extended version of the RAREEVENTS procedure that uses the limits from a previous procedure is shown below:

```
proc rareevents data=feedings limits=feedlimits;
     chart time_between / dist=Weibull totpanels=1;
          readphases=all limitphases=all
          phaselegend phaselimits;
run;
```

In this code, a few options are added. Here, `limitphases` specifies the phases for which probability limits are read from the `feedlimits` data set. The option `phaselegend` displays phase labels in a legend across the top of the chart, and `phaselegend` adds labels to the probability limits for each phase [SAS/QC, 2018, ch. 17].

3.3 Radar Chart

A radar chart shows the relative frequency of data measures. On a radar chart, the statistics are displayed along spokes that radiate from the center of the chart [SAS/GRAPH, 2016, ch. 39]. In SAS a radar chart is constructed by use of the GRADER procedure:

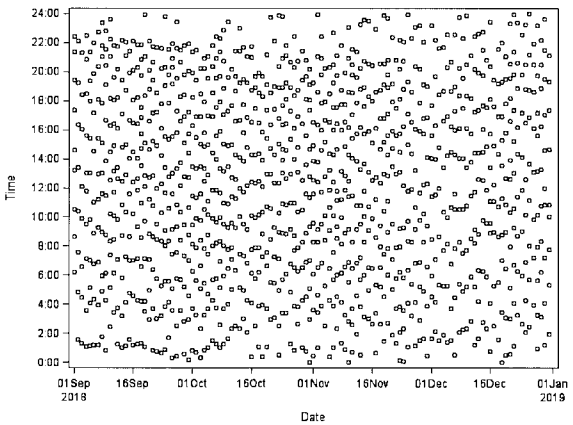
```
proc gradar data=feedings;
      chart time / freq=count;
run; quit;
```

In this code, `time` is the chart variable which defines the categories of data to be charted. The `freq=count` option specifies that the values weights from `count` gives the contribution of each observation in the computation of the chart statistic.

4 Results

When conducting an immediate visual analysis of a baby's meals on a four-months period, it can be very difficult to see any kind of pattern, trend or deviation. Figure 2 shows a scatter plot of the date and time for the baby's 1,229 meals. Each little circle represents a unique meal. There might be a slight tendency to a lower number of feedings between 22:00 and 24:00. However, it is not completely clear in this plot.

Figure 2: Date and Time for Feedings



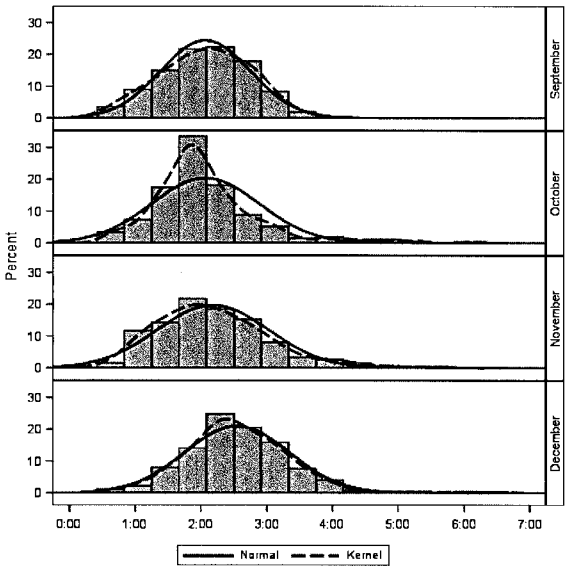
Source: Output from PROC SGPLOT, SAS® 9.4.

Further analyses are needed, to try to get a deeper understanding of the baby's eating habits.

4.1 Distribution of Feedings

A rather simple way of getting insights into data of the baby's eating behavior is by conducting an explorative analysis. This provides a visual overview of the distribution of time between feedings, shown in Figure 3.

Figure 3: Distribution of Time Between Feedings



Source: Output from PROC UNIVARIATE, SAS® 9.4.

Generally, there is not huge differences between the four months. One think to note is that the distributions seems to shift a bit to the right, for every month that goes. Further the distribution for October has a spike a little under 2 hours on time between feedings. Looking at the fitted normal and kernel curves, non seems to fit perfectly. When performing a statistical test for whether the four distributions are normally distributed, it is only for September, the hypothesis cannot be rejected.

To get a better picture of what is going on in Figure 3, descriptive statistics are calculated and displayed in Table 1.

Table 1: Descriptive Statistics of Time Between Feedings

Month	N Obs	Min	Max	Mean	Median	Std Dev
September	313	0:32:00	4:00:00	2:03:58	2:05:00	0:40:50
October	331	0:35:00	6:11:00	2:03:55	1:56:00	0:48:53
November	308	0:33:00	5:36:00	2:10:53	2:05:30	0:50:40
December	277	0:38:00	6:30:00	2:32:53	2:29:00	0:47:30
All	1229	0:32:00	6:30:00	2:12:12	2:08:00	0:48:27

Source: Output from PROC SGPLOT with SCATTER statement, SAS® 9.4.

The second column in Table 1 displays the number of observations for each month, corresponding to the total number of times, the baby has eaten. October was the months of most meals, whereas December had the lowest number. In the next two columns, the shortest and longest times between two consecutive feedings are given. Here, it is seen, that September had the shortest time for both, with 32 minutes being the shortest time observed between two meals. Cf. Section 2.1, all feedings with less than 30 minutes to the subsequent, have been deleted. In contrast December had the longest time between two meals, the longest being 6 hours and 30 minutes. Comparing both the mean and median times between feedings for the first months with the last, it is seen, that both statistics increase with almost 30 minutes. Only for September, the median is larger than the mean, which contrasts the three subsequent months. When the median is larger than the mean, the distribution of the data is said to be left skewed. The opposite applies when the mean is larger than the median, in this case, the data is right skewed. This also appears to some extent in Figure 3, where more observations with longer time between feedings are observed in the lower three histograms. Finally, the standard deviation, expressing how much the time between feedings differ from the mean, only differs with 10 minutes between the four months.

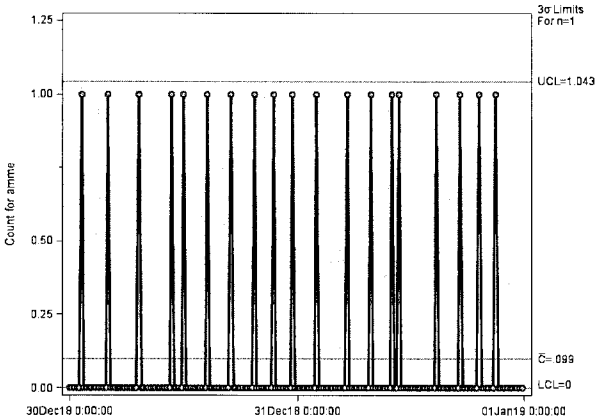
Both Figure 3 and Table 1 indicate that the baby needs food more often in the beginning of his life, and not nearly as often when getting older. December is the month, where the baby eats less frequently.

4.2 Variation of Feedings

To get an understanding of whether and how the number of feedings varies across time, a control chart is produced. The entire data set consists of 11,712 observations, representing a time series with 15-minute intervals. The number of feedings that occur at each 15-minute interval is used to construct a *c* chart, cf. Section 3.1. Figure 4 shows feedings over a two days period. It is seen, that within each 15-minute interval, there can either be one or no

feedings. Hence, all values lie within the LCL and UPL, and there is no sign of unusual variation. Therefore, this chart does not provide any interesting insights. If the intervals are changed to 1-minute intervals, the UCL would lie just below 1, which would result in all feedings signaling unusual variation. This would also not provide any interesting insights. In this context, the traditional control chart approach is of little use.

Figure 4: *c* Chart for Feedings per 15 Minutes

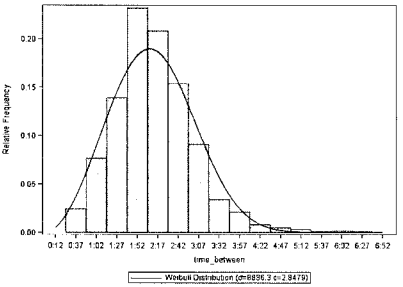


Source: Output from PROC SHEWHART, SAS® 9.4.

Instead of considering the number of feedings per time event, the time between each meal is considered. In this way the dataset is reduced from 11,712 to 1,229 observations. When using the time between feedings, it is possible to compute two limits, which can be used to indicate whether an observation is rare or not. Initially, a comparison plot is created, to figure out whether data distribution is appropriate for the theoretical distribution. Usually, the geometric or exponential distribution is used. As seen in Figure A8 in Appendix, these distributions do not match the data. Instead the Weibull distribution is used, as this has a better fit to data, cf. Figure 5.[1]

[1] Even though the fit is better, it is still not perfect. The goodness-of-fit tests indicate that the Weibull distribution is not accepted to fit data on a 95pct. confidence level, cf. Table A2 in Appendix.

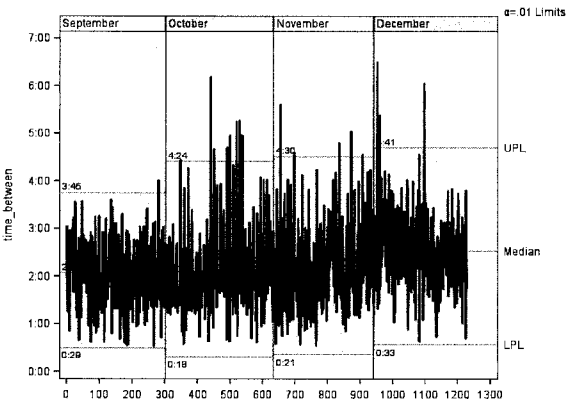
Figure 5: Distribution of Time Between Feedings, Weibull



Source: Output from PROC RAREEVENTS, SAS® 9.4.

The produced rare events chart for time between feedings is shown in Figure 6. This is an advanced rare events chart, as limits are calculated separately for each of the four months. A simple rare events chart is included in Figure A9 in Appendix. With the advanced chart, it is possible to compare the development in the variation of data. Figure 6 shows, that there are no values less than the LPLs, which indicates that the baby does not eat more frequent. In contrast, a few observations are greater than the UPLs, indicating that the time between feedings increases. This can definitely be considered as an improvement, as longer times between feedings leads to more time to do other things.

Figure 6: Rare Events Chart for Time Between Feedings



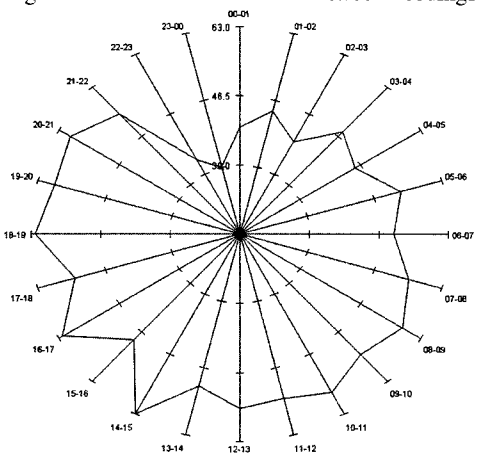
Source: Output from PROC RAREEVENTS, SAS® 9.4.
Note: The median for the four months are 2:04, 2:00, 2:08 and 2.31.

The median in Figure 6 does not change a lot, until the last month. Similarly, the LPLs are quite constant throughout the entire period. In addition, most feedings are of the same length, somewhere between 2 and 3 hours. What does change are the UPLs, which is equivalent to an increase in longer time between feedings. This can also be shown, by continuing the UPL and LPL from September, and using these for all events in the chart. Hence, more points would exceed the UPL. This is a strong indication, that the data from the first month comes from a different distribution than the rest. The process changes over time, as the baby learns to survive longer without food.

4.3 Frequency of Feedings

So far, it has been showed that there is a tendency, that the time between feedings gets longer, as each day goes. To get more understanding of which time periods during the day, the baby usually eats, a radar chart is generated.

Figure 7: Radar Chart of Time Between Feedings



Source: Output from PROC GRADAR, SAS® 9.4.

In Figure 7, a radar chart for all four months is given. The chart shows that the fewest number of feedings have occurred between 11pm and 12pm. Only 30 times has the baby eaten in this period. In contrast, the most common time to eat is between 2pm and 3pm, with a total of 63 meal. On this chart, compared to the scatter plot in Figure 2, it is easier to see, that there is less activity in the night times from 22pm until 5am. This is good news to a mom, who appreciates a good night's sleep.

5 Conclusion

Babies are unpredictable. Every time you think, you have understood what they are up to, they change everything. This can be extremely frustrating for parents, as it is difficult to make plans, as you never know what the baby is up to.

Considering data from this article, as if they were gathered from an average baby, then most new mothers do not need to worry too much. The results indicate that as the baby gets older, the number of daily feedings decrease, as the time between consecutive feedings increase. Further, there is tendency to less activity, when it comes to the baby wanting food. So even though everything might look very bleak, hope is ahead.

And of course, one must not forget, that even though parents give up a lot of things, such as their freedom, the little ones are the most wonderful things in the world. Parenting life is a blessing.

References

[NDBN, 2018] National Diaper Bank Network (2018). *Diaper Facts*. From http://nationaldiaperbanknetwork.org/ (accessed 31.12.2018)

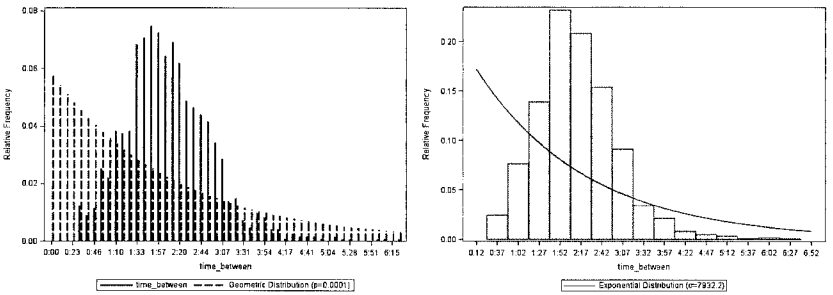
[Ransdell, 2016] Ransdell, B. (2016). *Improving Health Care Quality with the RAREEVENTS Procedure*. Cary, NC: SAS Institute Inc.

[SAS/GRAPH, 2016] SAS Institute Inc. 2016. *SAS/GRAPH® 9.4: Reference, Fifth Edition*. Cary, NC: SAS Institute Inc.

[SAS/QC, 2018] SAS Institute Inc. 2018. *SAS/QC® 15.1 User's Guide*. Cary, NC: SAS Institute Inc.

Appendix

Figure A8: Distribution of Time Between Feedings, Geometric and Exponential



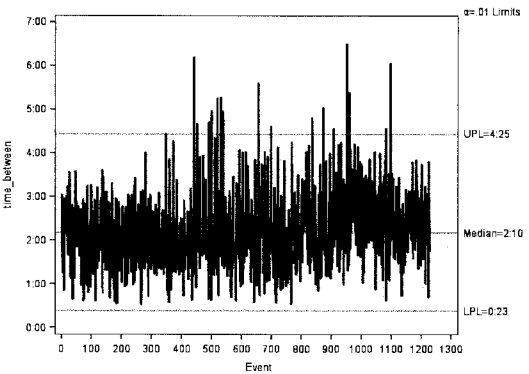
Source: Output from PROC RAREEVENTS, SAS® 9.4.

Table A2: Goodness-of-Fit Tests for Weibull Distribution

Test	Statistic		p Value	
Cramer-von Mises	W-Sq	0.94505809	Pr > W-Sq	<0.010
Anderson-Darling	A-Sq	6.53107142	Pr > A-Sq	<0.010

Source: Output from PROC RAREEVENTS, SAS® 9.4.

Figure A9: Simple Rare Events Chart for Time Between Feedings



Source: Output from PROC RAREEVENTS, SAS® 9.4.

The cumulative distribution function of a Weibull distribution with scale parameter σ, shape parameter c and threshold parameter θ is given in Equation A1.

$$\text{cdf}(x) = 1 - \exp\left(-\left(\frac{(x-\theta)}{\sigma}\right)^{c}\right) \tag{A1}$$

The associated probability limits are given in Equations A2.

$$\begin{aligned} \text{LPL} &= \theta + \sigma(-\ln(1-\alpha_{LPL}))^{1/c} \\ \text{median} &= \theta + \sigma(\ln(2))^{1/c} \\ \text{UPL} &= \theta + \sigma(-\ln(\alpha_{UPL}))^{1/c} \end{aligned} \tag{A2}$$

Mikkeller – a third attempt

Lisbeth la Cour, Dep. of Economics, CBS
Anders Milhøj, Dep. of Economics, KU
Ravi Vatrapu, Dep. of Digitalization, CBS

1. Introduction.

The present study is a continuation of the analyses presented in Buus Lassen et. al. (2017a) and la Cour et al (2018) in that it still focuses on how to model and predict series of interest to the management of a private firm using social media data. Our case company is as before: Mikkeller (a microbrewery). As in la Cour et al (2018), we focus exclusively on investigating predictive power of Facebook data (FB) with respect to sales. As mentioned in the paper above: "The main advantage of using social media data as predictors lies in the speed with which such data can be extracted and employed in the forecasting process. Once a firm has learned how to collect and pre-process their social media data, the information is available almost in real time and this implies that such data in combination with a good predictive model will provide a very useful tool for the management of the firm."

In the present study we still focus in the daily frequency of observations. Compared to the 2018-study mentioned above we add more deterministic dummies to capture calendar effects and holidays. Also we will look into specific non-linear transformations of the FB variables to see if that would increase the role of such variables in the models. The sales data comes from Mikkeller's bar in Viktoriagade (Vik) and FB data comes both from the pages of this bar but also from the Mikkeller Headquarters (HQ). We will apply two different times series techniques and we will also extend our focus to also include not just the sales series but also the volatility of the sales series.

2. Briefly on the existing literature.

The idea of using social media data as predictors for e.g. company sales is not new. When it comes to model building, various experiments have been conducted and a summary of around 40 articles covering the time period 2005 – 2015 can be found in Buus Lassen et al (2017b). For the present purpose the most interesting observations from these studies are that 1) almost 50% of the studies use some kind of regression model as their predictive model, 2) the range of social data types studied seem to cover Facebook, Twitter, Google Trends, Instagram, Tumblr, blogs and Youtube. In the literature, volatility modelling has been directed towards financial markets applications and has not been done for company sales series.

Theoretically, the argument for considering social data activity as predictors for sales obtains support from e.g. the AIDA model mentioned in Buus Lassen et al (2014). AIDA means *Awareness, Interest, Desire and Action* and refers to stages in a sales process. If social media data help increase the attention or can be considered a proxy for attention towards a product, then it may also affect the final decision about buying. It is the general perception that more attention will increase sales even if the attention is negative. The exact time pattern of social media reactions may not be of importance only for the sales series itself but also for the volatility of sales.

3. The data and methodology.

As stated in la Cour et al (2017): "As in our previous studies, we use data from Mikkeller's accounting system combined with Facebook data. In this analysis we have obtained daily sales data from a number of Mikkeller bars in the Copenhagen area: Viktoriagade, Stefansgade and Torvehallerne (the latter is also a Bottle Shop). The data from the bars are quite ideal for our purpose as they will relate directly to consumption of the product and therefore simplifies the way that we think about the lag patterns in the data. The time span of the study has been limited by our access to historical sales data and covers 1 January 2015 – 30 September 2017. In total we have 1004 observations. In order to perform an out-of-sample forecasting exercise we have held back 3 months of sales data as a tests sample while we select and estimate our model based on the remaining around 900 observations.

Prior to analysis we index the sales data such that the mean is restricted to 1234 and the standard deviation to 12. Such transformations do not affect the significance our results later in the modeling process. The Facebook data comes from the overall HQ Mikkeller FB page

https://www.facebook.com/mikkeller/
https://www.facebook.com/events/)

and from the FB pages of the chosen bars

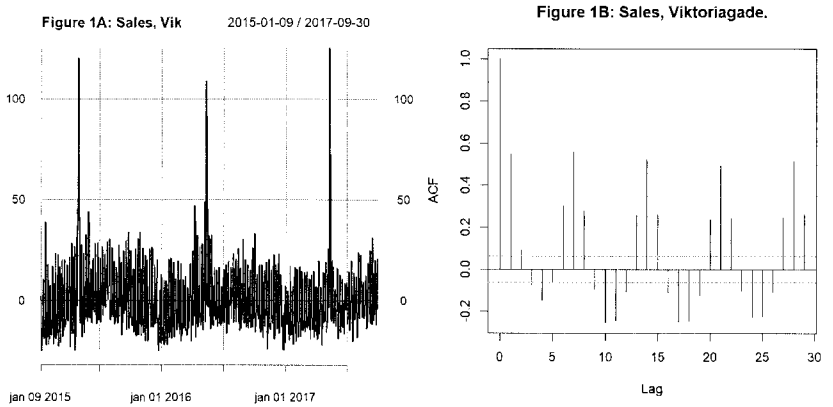
https://www.facebook.com/mikkellerbarvik/,
https://www.facebook.com/MikkellerandFriendsBottleShop/,
https://www.facebook.com/mikkellerandfriends/).

Using the Sodato software developed by Ravi Vatrapu and his group, see Hussain & Vatrapu (2014), we collect information from the selected FB pages and we create variables for e.g. total likes of the posts on a specific date. The 'posts' that we count are from the page administrators and therefore can be considered information from the firm to its customers."

3.1 Pre-processing methodology

Our first considerations when it comes to data preparatory work concerns whether to use simple transformations of the series or just the raw series themselves. As the values of sales are quite low on certain dates it does seem like a disadvantage rather than an advantage to use a log-transformation. As the sample period contains three dates (1/1 2015, 1/1 2016 and 1/1 2017) where the bar is closed we have some 'artificial' zeros in the series. If we just dummy out these dates during the regressions analysis we will end up with some artificially high R squares, so we decided to start out by regression the indexed sales series on the dummies for the closing dates. We save the residuals from this regression and use these as the left-hand-side variables in our regression models (or as our main variables of interest).

We show graphs of our main series of interest and its ACF in figures 1A and 1B.



The sales series is characterized by seasonal fluctuations, no specific trending behavior (or only a very weak decline) and some large outliers that are found during the annual Mikkeller Beer Celebration Copenhagen (MBCC) week. We see some significant autocorrelations around the first few lags and also around the seasonal frequency of a week (lag 7). The autocorrelations are not large enough to make us suspect non-stationarity of the series. The reason why some values are negative in Figure 1A is that we show the series after our initial preparatory transformations to anonymize the data and to prevent artificially high R squares due to closing days of the bar. We do not pre-process the FB data.

When studying the predictive properties of FB data we follow two different strategies as we also did in Buus Lassen et al (2017) and la Cour et al (2018). One direction of analysis is based on a split of the data into components of trends, seasonal terms and irregular parts. Such models are known in the literature as Unobserved Component Models (UCM). Our second direction of research builds upon regression models.

3.2 Unobserved Component Models

We use the same modeling strategy as in Buus Lassen et al (2017) and therefore start out by employing an unobserved component (UCM) model. An UCM decomposes the observed series y_t into a sum of many components, as for instance

$$y_t = \mu_t + \varepsilon_t$$

$$\mu_t = \mu_{t-1} + \eta_t$$

Here the series μ_t is understood as the level of the series; but this level is unobserved. Only the series y_t which is affected by some noise or irregularities is observed. This noise series, ε_t, could in technical applications be measuring errors.

This basic formulation could be extended by trends and seasonality, and various forms for introducing autocorrelation in the model formulation also exist. A trend component is insignifikant for the sales series. A seasonal component for the day of the week effect is defined in a way so it does not affect the level component:

$$S_t = -(S_{t-1} + ... + S_{t-6}) + \zeta_t$$

The yearly variation is modelled by a cycle component; in fact only a single cosine for approximately 365 days is needed as the yearly variation is very regular. Moreover, all calendar dummies for e.g. Easter are included in the model as regression terms and also a first lag autoregressive term is needed because of autocorrelation.

In total these ideas lead to the model:

$$y_t = \mu_t + S_t + \text{cycle} + \text{calendar dummies} + \varepsilon_t + \varphi_1 \varepsilon_{t-1}.$$

All remainder terms, ε_t, η_t, and ζ_t, are assumed to be mutually independent white noise series. Their variances could be estimated; the larger this component variance the more volatile the component. But it is also possible to fix this variance to the value zero which gives a constant component, e.g. a model with fixed seasonal dummies is found if $\text{var}(\zeta_t) = 0$.

In the first round the parameters of this model are estimated by the Kalman filter extended by a further smoothing estimation algorithm. From the estimation results all significant calendar dummies are identified. The idea is then to simply remove observations for all these dates instead of keeping the dummies in the final model. In this way these observations of the sales series are set to missing instead of letting them be dummied out.

In the last estimation round the unobserved component model is estimated once more for the series with all observations with extraordinary calendar effects set to missing - of course the calendar dummies are now excluded.

The final result is a series of residuals defined as one day ahead forecast errors. These residuals represent sales number corrected for seasonal effects and autocorrelations. Note that these residuals are not defined for the days during the beer festival and days with calendar effects.

3.3 The regression models

Secondly, we use dynamic regression models (or ADL models). With daily data we have a rich seasonal structure and even though we only have a sample period covering less than three years we have enough observations to model the seasonality by deterministic dummies. Using lags of both the dependent variable and the independent variables is also possible and we will do both. As scraping Facebook pages and pre-processing of the data is time-consuming and costly we will start by building a benchmark model for sales based on only a trend variable, deterministic seasonal dummies and lagged values of sales itself.

The model we take as the starting point for the benchmark model is:

(1) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_h y_{t-h}$ + seasonal dummies + trend + ε_t $\quad t = 1, .., T$ (GUM1)

where y is sales. We also include dummies for calendar and holiday effects[1]. This makes the model suitable for at least one step ahead predictions out-of-sample. We use up to eight lags to cover a same-day-of-the-week effect and also an interaction of short run and day-of-week effects. The error term, ε_t, is assumed to fulfill the standard assumptions for OLS estimation.

We call this model our benchmark GUM and we use this large model as a starting point for model selection for a more parsimonious benchmark model. The terminology GUM (General Unrestricted Model) comes from the Autometrics/GETS model selection sphere. This model selection strategy was developed by Hendry and co-authors and made available to researchers in the Autometrics tool in OxMetrics but it is also now available in the GETS function in R, see e.g. Doornik & Hendry (2014), Hendry & Pretis (2013) and Pretis et al (2018). We will use the latter as it also allows modelling of the conditional variance.

It is important that the 'GUM' model has statistically well specified error terms. We will check the usual assumptions of no autocorrelation, no ARCH and normality by tests like Ljung-Box, ARCH and Jarque-Bera tests. The first two tests are considered more important than the last one. Sometimes we will also have to close fown the

[1] The full list of calendar effects, we consider, are: Easter, Whitsun, Ascension, St. Hans, May1, 'St. Bededag', 23/12, 24/12, 25/12, 26/12, 31/12 and the Danish constitution day, 5/6. Also we include dummies for day-of-the week (Monday being the base category) and month of the year (January being the base category). We construct sets of dummies for each of these events.

ARCH test initially to get a starting point as in equation (1). To briefly sketch the idea of this back-ward selection procedure, it starts from a well-specified GUM. The reduction is tried by sequentially dropping variables that are insignificant but at the same time ensuring that the reduced model is still statistically well-specified. A path of reduction is followed for each insignificant variable in the original GUM. This may lead to more than one suggested simplified model after following all such paths and in the end all of these reduced models are compared to each other based on a selection information criterion to find the 'final' parsimonious model. The researcher can influence many of the criteria and tests that are used during this process. For a given GUM the outcome may be that no statistical reduction is valid. This will happen if none of the reduced model candidates fulfill the error assumptions selected for the GETS procedure. For selection we choose to use 'Whites' standard errors as we believe that the variance may be affected by certain FB regressors. As long as these are not included in the model we rely on such robust standard errors.

Next, we need to specify GUM's that include various sets of FB variables. In all cases we focus on only lagged FB variables as the model is constructed for predictive purposes. We will work with a few selected GUMs for inclusion of FB variables:

(2) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_h y_{t-h} + \beta_1 \Sigma x1_{t-i} + \ldots + \beta_k \Sigma x k_{t-i} +$
$+$ seasonal dummies + trend + $\varepsilon_t \quad t = 1, .., T$ (GUM2)

the x's are the four lagged FB measures and the sub scripts, i, runs from 1 to 8. We consider this model as a simple, linear model to get started when introducing FB variables.

Next we consider two GUM model versions where we in addition to the linear FB terms include either quadratic terms or dummies to identify if a FB value lies above the 90 quantile (a threshold effect idea). We also include up to 8 lags of these non-linear FB terms:

(3) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_h y_{t-h} + \beta_1 \Sigma x1_{t-i} + \ldots + \beta_k \Sigma x k_{t-i}$ + lagged quadratic FB terms + seasonal dummies + trend + $\varepsilon_t \quad t = 1, .., T$ (GUM3)

(4) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_h y_{t-h} + \beta_1 \Sigma x1_{t-i} + \ldots + \beta_k \Sigma x k_{t-i}$ + lagged threshold FB terms + seasonal dummies + trend + $\varepsilon_t \quad t = 1, .., T$ (GUM4)

Notice that by including dummies for the MBCC events we would for all types of models increase the explanatory power considerably. (These results are available from the authors upon request). However, such a model would not be suitable for predictive purposes as the coefficients of these dummies vary from year to year. Hence we cannot rely on coefficients from historical episodes to predict future sales during such event weeks.

Finally, we consider models that contain equations for both the conditional mean and the conditional variance in a kind of ARCH spirit. In our case the main interest is in

adding FB variables to the equation for the conditional variance to investigate if the volatility in sales is affected by social media activities.

We start out by the GUMs already defined by equations (1) to (4). We then investigate whether adding FB variables to the conditional variance equation[2] helps improve the models.

The conditional variance equations are for the log of the variance (our left-hand-side variable) to ensure a positive conditional variance. The 'i' (i = 1,...,4) in the equation number signals that we specify conditional variance equation that corresponds to the equations used for the conditional means. We do, however, not include the seasonal and calendar terms in the equations for the conditional variances.

(i) $\log(\sigma_t^2) = \alpha_0 + \alpha_1\varepsilon_{t-1}^2 + \ldots + \alpha_h\varepsilon_{t-h}^2 + \delta_1\Sigma x1_{t-i} + \ldots + \delta_k \Sigma xk_{t-i}$ + maybe more lagged FB terms

(As in standard ARCH-type models we now assume that the error terms of the conditional mean equation can be specified as: $\varepsilon_{t-1} = \sigma_t z_t$ where $z_t \sim N(0,1)$).

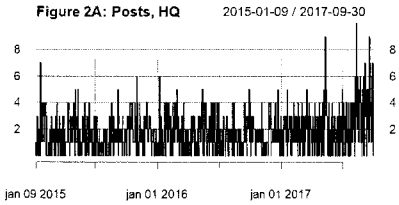
4. Descriptive statistics.

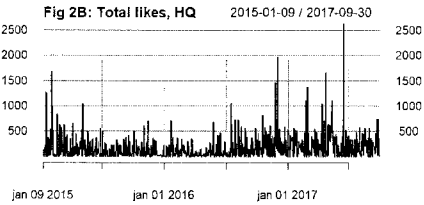
We have already shown graphs of the indexed sales series and its ACF. To get a first impression of the some of the data from Facebook, Figures 2A – 2D show the number of likes and the number of posts by the administrator for the HQ page and for the Viktoriagade page. These graphs can also be found in la Cour et al (2018). While none of the series seem to follow the pattern of the sales series very closely they seem to correlate pairwise (Viktoriagade – Viktoriagade and HQ – HQ). Also the number of Likes for HQ are - not surprisingly - in general larger than for Viktoriagade. Notice that the activity for Viktoriagade show a decline in 2017 compared to the other years (Mikkeller has no specific explanation to that).

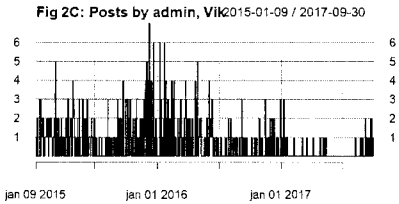
Instead of visually trying to assess the presence of relationships between sales and the FB variables we will try to employ automatic model selection routines to backwards select from small amount of benchmark or other models. For more details on this modelling strategy, confront section 3.3.

Table 1 shows simple descriptive statistics for the variables we have been investigating so far. The numbers of the mean and standard deviation for sales reflect our standardization and pre-modelling of closing days. Not surprisingly we see both more posts and also more reactions to the HQ FB activity.

[2] We actually set up the model for the conditional variance as a model for the residual from the first stage conditional mean model. In the end we report results for a model for a reduced form conditional mean and variance estimated in one step by the ARX function from the GETS package.







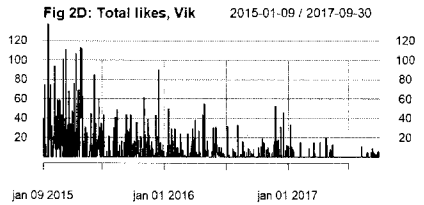


Table 1: Descriptive summary statistics.

Variable	N	Median	Mean	Std Dev	Minimum	Maximum	N Miss
'Sales'	996	-4.50	0.07	16.98	-24.92	125.40	0
Posts by Admin (VIK)	996	0	0.57	0.95	0	7	0
Posts by Admin (HQ)	996	2.00	1.74	1.48	0	10	0
Likes by Viktoriagade	996	0	6.68	15.71	0	137	0
Likes by HQ	996	91.50	168.71	231.37	0	2626	0

Note: 'Sales' are the anonymized, preprocessed numbers.

5. Unobserved components models for the sales series

In the first step of the model fitting in total 43 observed daily sales were set to missing because of the beer festival significant calendar effects; this leaves us with 861 observations. The resulting model in the second step is:

$y_t = \mu_{t-1} + \text{cycle} + S_t + \varepsilon_t + \varphi\varepsilon_{t-1}$

The variance in the level component is fixed to zero, meaning that the level is constant. The variance for the seasonal factor is borderline significant, p =4.1%, as the spread between the day of the week with the smallest sales and the day of the week

with the largest sales are reduced by approximately 18% in the data period. The autoregressive parameter is $\varphi = 0.43$.

The resulting series of residuals from the UCM model in the previous section has variance 5.94^2. In this section we try to include the FB lagged FB variables in order to reduce this variance, i.e. to see if we could forecast these residuals by previous posts or likes. If this is possible we could optimally affect future sales by FB-activities.

All attempts by inclusion of lag one to eight of the social media data only gave spurious significances; some of the even with an unexpected negative coefficient. But the overall picture was that the coefficients were mostly positive. We then concluded that it is impossible to claim that sales depend only on the total FB-posts from the bar say four days ago and not on the number of posts all other days the previous eight days. It was then decided to consider the sum of likes or posts the previous eight days as the input variable. This means that we expect the sales to depend on the FB activity the previous eight days, but we cannot tell what the precise time lag is.

The numbers of posts (or likes) for eight days could be zero and furthermore they could very large, so it was decided to transform by the function log(x+1). The '+1' is because of the zeroes.

If we systematically apply all possible combinations, we find that only three of the potential variables are significant - but, however, only when a model with only one exogenous variable is applied. When two exogenous variables are applied in the same model, only one of them is significant. The table summarizes the estimation results for these three models.

Exogenous variable	Coefficient	p-value	R-squared
Likes for the local bar	0.418	0.0012	0.0123
Posts from the local bar	0.773	0.0025	0.0107
Likes for the company	- 0.603	0.0416	0.0049

Note: The variables are the total number of likes/posts from eight days before to just one day before. These numbers are then transformed by log(1+x).

The least significant exogenous variable, likes for the company, however turns out to have an unexpected negative sign; so we regard this as a type I error.

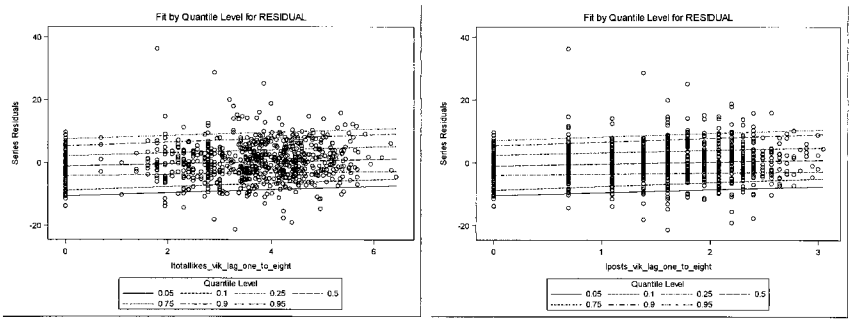
The next three figures present results which could help to understand how the exogenous variables affect the distribution of the residual series. The quantile plots from quantile regressions have been made for the two exogenous variables of lagged numbers of lags from the local bar in Viktoriagade. The fitted regression lines are for

the 0.05, 0.10, 0.25, 0.50, 0.75, 0.90 and 0.95 quantiles. The reported confidence intervals for around half of the regression coefficients do not include zero, so even if these facts in no way are a result of pure statistical testing at least some impact of the variables are seen. On the other hand, the impact is judge from the plots very small. As the lines look fairly parallel, the actual shape of the distribution of the residuals is independent of the exogenous.

However, regression models tell that the log transformed squared residual to some extend are affected by two of the exogenous variables. The total likes for the local bar tend to increase this variance measure while the total posts from HQ decreases the variance - these effects are however small.

The conclusion is then that the effect of the exogenous variables mainly tells something about the position parameter in the distribution of the residuals - and that this impact is very, very small.

Figure 3: Quantile plots for regression of the residuals on the log transformed, lagged numbers of total likes respectively posts from the local bar.



6. Results of the regression models

6.1 Estimation results, regression models

Estimation results for a selected range of regression models are shown in Table 2. The column numbers correspond to the equations (1) – (4) as mentioned earlier. These are the GUMs used for the model reduction exercises. Table 2 shows the results of the final reduced models that resulted from the GETS procedure i.e. when also the conditional variance has been modelled.

To save space we will just comment on the results of most of the dummies without actually showing the coefficients. For all 24/12 and 31/12 dummies the coefficients are

negative and maintained in the models. The other Xmas dummies were not consistently picked by GETS. In all models the monthly dummies for April and May were always picked while other moenthly dummies were picked in some models but not in all. In all models the coefficient for May is the largest probably capturing some of the effect of the MBCC events. For the day-of-week dummies (all of them picked in all models) the coefficients for Tuesdays and until Saturdays are always positive indicating that Mondays (the base category) have lower sales than these other week-days. Fridays have always the largest coefficients followed by Saturdays and Thursdays (maybe not very surprising). Sundays have significantly lower sales even than Mondays. We also have selected holiday dummies left in all the models but not always the same ones. Coefficients of these are available upon request.

In Table 2 we only see few significant FB variables and as mentioned before none of the quadratic terms were included in the final specifications. Surprisingly some of the remaining FB variables show up with negative coefficients in the models. This predominantly happens for variables of the 'Likes' type. As we do not in the present setting know whether a 'Like' comes after a positive or a negative 'post' it is difficult to say much more about this. Column (4) indicates that the threshold transform FB variable seem to play a somewhat larger role than the original linear terms. Also notice that the coefficient of the threshold variable for posts by Viktoriagade Bar itself has the largest and most significant coefficients of the selected FB variables. Still the increase in R^2 compared to the benchmark model (1) is limited or non-existing and the overall value of including FB variables like those of the present study is questionable. At least when it comes to the conditional mean. A suggestion for further research could be to go back to the actual text of the HQ and vik posts and apply text analytics to their content to better classify the single posts in accordance to their purpose/content. Especially for the posts in the weeks up to the MBCC event some text analytics might help identify proper marketing actions and may help suggest how to create better explanatory variables, Such an exercise is, however, outside the scope of this paper.

Table 2: Results from the final conditional mean regressions

	Dependent variable:			
	SALES			
	(1)	(2)	(3)	(4)
sales1	0.602***	0.593***	0.615***	0.610***
	(0.030)	(0.029)	(0.029)	(0.029)
sales3	0.126***		0.150***	
	(0.035)		(0.034)	
sales4	-0.125***	-0.115***	-0.115***	-0.086***
	(0.037)	(0.032)	(0.032)	(0.029)
sales5	-0.052	-0.092***		
	(0.037)	(0.033)		
sales6	0.061*	0.064*		
	(0.034)	(0.033)		
sales7				0.058*
				(0.033)
sales8	-0.062**	-0.066**		-0.096***
	(0.029)	(0.029)		(0.033)
totallikes_HQ_lag7		-0.003**	-0.003**	-0.003*
		(0.001)	(0.001)	(0.001)
posts_HQ_lag2		0.668**		
		(0.329)		
posts_vik_lag3		0.668*		-1.173**
		(0.329)		(0.551)
totallikes_HQ_90_lag1				-1.797*
				(1.059)
posts_vik_90_lag3				6.526***
				(1.585)
Observations	904	904	904	904
R2	0.737	0.733	0.729	0.737
LB AR(1) test	0.908	0.126	0.421	0.084
LB ARCH(9) test	0.033	0.006	0.926	0.032
JB test	0.000	0.000	0.000	0.000
Note:				*p<0.1; **p<0.05; ***p<0.01

Note 2: Standard errors in parentheses.
Note 3: A lot of dummies for day-of-week, months-of-year and also certain holidays are also kept in the model but for ease of exposition not displayed in Table 2. Also a constant term is present in all models.
Note 4: Standard F-tests for all models indicates strong significance.
Note 5: Misspecification tests for no AR(1) errors frail to reject the null of no autocorrelation at the 5% level. The ARCH(9)tests are weakly rejecting in three out of the four models.

6.2 Results of estimations that include FB variables in the conditional variance equation.

We now turn to the reporting of the results for the conditional variance to see if our FB variables seem to play a larger role when it comes to the uncertainty of the sales behavior. The results of the final models are shown in Tabel 3.

Table 3: Results from the final conditional variance regressions

Dependent variable:	log(sigma-t-squared)			
	(1)	(2)	(3)	(4)
Constant	3.900***	2.703***	2.861***	2.738***
	(0.161)	(0.299)	(0.318)	(0.297)
arch1	0.051	-0.051		
	(0.033)	(0.035)		
arch2	0.080**	-0.0080		
	(0.033)	(0.035)		
arch3	-0.004***	-0.013***		-0.070**
	(0.033)	(0.035)		(0.034)
arch6	0.032	0.009		
	(0.033)	(0.035)		
arch7				-0.074**
				(0.033)
arch8				-0.087***
				(0.033)
logEwma(7)		0.319***	0.353***	0.395***
		(0.102)	(0.080)	(0.088)
Totallikes_HQ_lag7			0.002***	0.001**
			(0.001)	(0.000)
Posts_HQ_lag1		-0.152**		
		(0.059)		
Posts_HQ_lag6			0.287**	
			(0.134)	
Totallikes_vik_lag1		-0.004		
		(0.006)		
Totallikes_vik_lag2			0.010**	0.010**
			(0.005)	(0.005)
Totallikes_vik_lag4		0.015***		
		(0.005)		
posts_vik_lag1		0.075		
		(0.099)		
posts_vik_lag4		0.050		
		(0.080)		
Totallikes_HQ2_lag6			0.000**	
			(0.000)	
Totallikes_HQ2_lag7			0.000**	
			(0.134)	
Posts_HQ2_lag6			0.000***	
			(0.000)	
totallikes_HQ_90_lag7				-1.035***
				(0.378)
posts_HQ_90_lag7				-0.394
				(0.273)

Note: *p<0.1; **p<0.05; ***p<0.01
Note 2: Standard errors in parentheses.
Note 3: We do not include the seasonal dummies in the equation for the conditional variance.

From table 3 we see that the FB variables seem to pop up in a significant manner but the pattern is not very systematic. In one model some sub set of lags is significant and in other models other lags are significant. One general observation is that lags around the weekly frequency do seem to have some influence. The coefficients of the quadratic terms although significant are very small and probably not economically significant. The coefficient of the Totallikes HQ threshold FB variable is significant and with a negative sign.

7. Future analysis

First of all we have noticed that using GETS more or less automatically may not lead to reduced models that necessarily contain interpretable terms. It may be worthwhile to spend more time on getting to reduced models that contain convincing terms before forecasting is considered. However GETS provides a very good starting point.

In the present study we focused entirely on in-sample estimation results and the next natural step will be a comparison of the models when it comes to an out-of-sample forecasting exercise. An external validation can also be performed by e.g. using the models from the Viktoriagade bar on data for some of the other Mikkeller bars. We will also need to consider in more detail how to model the MBCC events as these still seem to lead to large outliers in the models.

Compared to previous results we have some but still quite weak indications of influence from FB variables on the sales in the Viktoriagade Mikkeller bar.

8. Summary and conclusion

In this paper we have pursued our idea of applying a preparatory ucm model to both regressors and regressand to determine a forecasting model for the monthly sales of the Danish microbrewery Mikkeller. Also we tried a more traditional strategy with lagged sales to model the autocorrelation of the in the sales series and a suite of dummy variables for deterministic outside factors.

We do in both model approaches find minor evidence for a role for FB data in the models. This statement is based on statistical significance considerations as inclusion of FB variables generally do not improve much R^2-wise compared to benchmark models that are based on only deterministic seasonal dummies. A natural extension of our analysis will fine-tune the regressions models and also perform an out-of-sample exercise for both model set-ups.

9. References

Buus Lassen, N., la Cour, L., Milhøj, A., Vatrupu, R. (2017a), 'Social media data as predictors of Mikkeller sales?' in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 71-86

Buus Lassen, N., la Cour, L., Vatrapu, R. (2017b), 'Predictive Analytics with Social Media data' in Sloan & Quan-Haase ed. *The SAGE Handbook of Social Media Research Methods*, Chapter 20, pp 328-341

Buus Lassen, N., Madsen, R. and Vatrapu, R. (2014). 'Predicting iPhone Sales from iPhone Tweets', Conference Paper, *2014 IEEE International Enterprise Distributed Object Computing Conference.*

Buus Lassen, N., Vatrapu, R., la Cour, L., Madsen, R. and Hussain, A.(2016), 'Towards a Theory of Social Data: Predictive Analytics in the Era of Big Social Data', in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 241-256

La Cour, L., Milhøj, A. , Vatrapu, R. and Buus Lassen, N. (2018). 'predicting the daily sales of Mikkeller bars using Facebook data', in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 125-141

Doornik & Hendry (2014). 'Statistical Model Selection with 'Big Data', *Department of Economics Discussion Paper Series*, University of Oxford, #735.

Hendry & Pretis (2013). 'Anthropogenic influences on atmospheric CO2', Chapter 12 in *Handbook of Energy and Climate*. Edward Elgar. Ed. By Fouquet. pp 287-326

Hussain A., Vatrapu R. (2014) Social Data Analytics Tool (SODATO). In: Tremblay M.C., VanderMeer D., Rothenberger M., Gupta A., Yoon V. (eds) Advancing the Impact of Design Science: Moving from Theory to Practice. DESRIST 2014. Lecture Notes in Computer Science, vol 8463. Springer, Cham

Pretis, Reade & Sucarrat (2018). 'Automated General-to-Specific (GETS) Modelling of the Mean and Variance of Regressions, and Indicator Saturation Methods for the Detection of Outliers and Structural Breaks'. In *Journal of Statistical Software*, vol 86, 3. Pp 1-44. Doi: 10.18637/jssv086.i03

TextClassify: Et ekspertsystems og menneskelige eksperters læring af semantik fra data

Birgitte Feiring(1), Adrian Hassler(1), Niels Jørgen Kjær(2*), Jonas Vago(2)
(1) Institut for Menneskerettigheder (2) Specialisterne ApS
(*) Kontakt: njk@specialisterne.com

Resumé:
Papiret beskriver den praktiske udarbejdelse af et ekspertsystem til brug for klassificering af tekstbidder til veldefinerede mål.

Kobling af 60.000 UPR anbefalinger til FN's 169 SDG delmål
FN har siden 2007 organiseret arbejde vedrørende menneskerettigheder i et Universal Periodic Review[UPR]. UPR har tilbagevendende sessioner som behandler menneskerettighedssituationen i de enkelte FN lande.
Et af resultaterne er rapporter med anbefalinger fra enkelte lande til landet der er under behandling. Dette land forventes så at adressere og forhåbentlig også implementere de enkelte anbefalinger. I løbet af de første 10 år er det blevet til ca. 60.000 anbefalinger. Dansk Institut for Menneskerettigheder, IMR, satte sig som opgave til at finde forbindelserne mellem disse anbefalinger og FN's Verdensmål for bæredygtig udvikling[SDG] for at vise, hvordan de kunne implementeres samtidigt. SDG'erne består af 17 hovedmål med 169 delmål, som hver især er tilknyttet målbare indikatorer.

IMR startede processen med små pilotprojekter på enkelte landes anbefalinger og enkelte SDG mål. Desuden blev der lavet manuelle forsøg på at lave logiske regler for koblingen mellem anbefalinger og SDG'er. Selv om alt dette fungerer rimeligt på små mængder data, viste det sig umuligt at skalere til ca. 60.000 koblinger i hele datasættet. De logiske regler er for uhomogene, endda upræcise for visse SDG delmål, og en ren manuel bestemmelse af koblingerne er ikke realistisk pga. de store mængder data.

Derfor iværksatte IMR i samarbejde med Specialisterne et samarbejde for at forsøge en BigData TextMining tilgang herefter kaldt TextClassify.

Samkøring af (fejlbehæftede) databaser.
De 60.000 anbefalinger findes som TekstBidder i FN dokumenter. Desuden findes der to databaser, [UHRI] og [UPR-INFO], som indeholder de enkelte TekstBidder.
Begge databaser har desværre udfordringer. Tekstbidderne fra UHRI indeholder irrelevant formateringsinformation, og tekster fra den første halvdel af perioden indeholder sporadiske forekomster af både uønskede mellemrum midt i orde ne og

manglendemellemrum. UPR-INFO databasen er næsten fejlfri, men indeholder kun teksterne på engelsk (UHRI byder på alle 6 FN-sprog) og har kun få ekstra metadata. Derfor valgte vi at sammenkøre de to registre, således at UPR-INFO TekstBidderne blev brugt til at validere flejbehæftede data fra UHRI. Metadata fra databaserne indeholder foruden logistisk information også en klassificering ud fra en menneskeretsvinkel med ca. 150 UHRI-klasser og 56 fra UPR-INFO.

Bag-of-Words naivt Bayesisk netværk på metadata.

På de sammenkørte data brugte vi først en simpel Bag-of-Words[BOW] model, implementeret som et naivt Bayesisk netværk med input baseret på de tilgængelige metadata fra de to databaser.
Resultaterne var tilfredsstillende men begrænsningerne er overvældende. Foruden at nogle metadata ikke var implementeret, så mangler sådan et setup robusthed. Det afhænger af at metadata er uforanderlige (dokumenterne strækker sig over mere end 10 år) og tilgængelige inden for et rimeligt tidsrum. Desværre er dette ikke tilfældet, og det gør vores behandling af data afhængig af andres systemer. For at omgå dette problem besluttede vi at TextClassify i fremtiden kører direkte på de originale TekstBidder. Disse uddrages med et simpelt UntaggedPDFtoStructuredTextSnippet script udarbejdet af Specialisterne som et andet mindre projekt.

VSO modellen
Vi nærstuderede det semantiske indhold af mange TekstBidder og deres SDG-kobling. Den egentlige SDG-koblende mening er begrænset til højst tre forskelligeTyper som er tilstede i forskellige Niveauer. Dette svarer nogenlunde til en lingvistisk Verbum-Subjekt-Objekt[VSO] model. Hver af de tre Typer kan optræde flere steder i TekstBidden og i flere Niveauer. Det samlede TypeNiveau bestemmes som det maksimale fundne Niveau for hvert af VSO-Typerne. Det er åbenlyst at behandle Niveauerne som diskrete værdier, og Niveauerne er naturligt begrænset til få værdier (0-3). Det samlede VSO udfaldsrum for den enkelte TekstBid er derfor begrænset til 64[0;3,0;3,0;3], hvor de individuelle Niveauer er strengt ordnede i forhold til kvaliteten af deres SDG-kobling.Verbet viser sig at være den dårligst separerende Type i modellen. Det skyldes at anbefalingerne næsten altid omhandler forbedringer af menneskerettighederne.

StamOrdbog og REGEX StåendeUdtryk
Forskellige bøjningsformer af ord har oftest præcist den samme SDG-kobling. Derfor samles alle bøjningsformer i en OrdBog med Ord->StamOrd. Dette relativt store arbejde var nødvendigt for at sikre en stor hyppighed af de enkelte StamOrd.
De betydende StamOrd er samlet i en EkspertOrdBog, som for hvert SDG delmål, indeholder StamOrdets Type+Niveau. Den kombinerede EkspertVærdi er så begrænset til "" uden værdi, "1-3" for Verbum, "4-6" for Subjekt, og "7-9" for Objekt.

Foruden StamOrd er det også muligt at inkludere StåendeUdtryk som EkspertOrd via REGEX-udtryk[REGEX]. Den nuværende EkspertOrdBog er således en sparsom matrix med værdier for 169 delmål gange 550 StamOrd + 250 REGEX-udtryk.

MahjongAlgoritmen

Når hver TekstBid er tildelt 3 VSO Niveauer, er opgaven at sortere udfaldsrummet med 64[0;3,0;3,0;3] felter efter renhed. Renheden er defineret som andelen af TekstBidder, der er GroundTruth[GroundTruth] for et givet delmål. Det viser sig at en Interior Point Barrier Method[IPBM] er velegnet til formålet.

Når de enkelte, V,S,O Niveauer antages strengt rangordenede i forhold til deres SDG-kobling vil:

a) antallet af mulige opdelinger i 2 underrum (barrierer) reduceres fra 2^64+1 til ca. 230.000 gyldige barrierer, som opfylder de individuelle rangordninger.
b) gyldige barrierer, mellem de mere og mindre rene felter, have entydige værdier, endda i alle projektioner.
c) sorteringen (lokalt) fremstå som et 1-dimensionalt problem.
d) sekvensen mellem en mere ren barriere og den næste (udvidet med et ekstra felt), mindre rene barriere, være veldefineret.
e) denne sekvensering være præ-programmerbar, således at hver tilladt barriere "peger" på få veldefinerede barrierer (med et ekstra felt) med mindre renhed.
f) behandlingen af tomme felter (uden TekstBidder) være triviel, da de er alle gyldige udvidelser, med tomme felter, mellem to barrierer med aftagende renhed.
g) udvidelsen hyppigt bestå af (først alle gyldige udvidelser med tomme felter plus) en udvidelse med det gyldige nabofelt, som har den højeste renhed.
h) mindre hyppigt, på grund af statistisk fluktuation, udvidelsen indeholde en "overset" ø med "unormal" høj renhed. Når denne ø inkluderes, kræver rangordningen, at et veldefineret sæt af andre felter "følger med".
i) der sjældent være mere end en enkelt overset ø. Disse situationer genkendes let og hurtigt, og sorteringen korrigeres derfor.

Ovenstående er kompliceret at beskrive i ord, men relativt let at programmere, og detaljerne fremvises (farvelagt!) i den mundtlige præsentation.

Bundlinjen i ovenstående er en robust sortering, der eksekveres på under 50 mikrosekunder(CPU), hvilket er en faktor 1000 gange hurtigere end en BruteForce metode. Denne voldsomme sorteringshastighed er grundlaget for at TextClasssify kan afprøve alle StamOrd, afprøve (næsten) alle kombinationer af to StamOrd, samt gøre det muligt at gøre indlæringen robust overfor overtræning ved anvendelse af Leave One Out Cross Validation[LOOCV], der bestemmer kvaliteten af klassificeringen på TekstBidder, uden at disse er brugt i sorteringen. Kvaliteten af klassificeringen

bestemmes af den samlede (summerede) effektivitet gange renhed. For at gøre kvalitetsændringerne lettere at forstå for ikke statistik-kyndige, oversættes dette (gange det totale antal) til antallet af MaskinLærteKoblinger.

EkspertSystemet (i brug fra 20180415)

TextClassify fungerer ud fra følgende filer (.csv.txt/.xlsx):

TextClassify som indeholder en søjle med tekstbidderne, en eller flere søjler med GroundTruth (alle delmål), og en søjle som specificerer om Tekstbidderne bruges i træningen eller kun evalueres på.

TextClassifyDictionary som indeholder Ord->StamOrd.

TextClassifyExpert som er EkspertOrdBogen beskrevet ovenfor.

TextClassifySteer som styrer, via nogle få parametre, hvordan optimeringen foregår.

Når EkspertSystemet kører, afprøves (optimeres) alle mulige forbedringer af EkspertVærdi, for alle StamOrd/REGEX, og for alle delmål.

Alle mulige udvidelser af TextClassifyExpert med StamOrd afprøves.

Alle ændringer (med høj kvalitet) anvendes efter godkendelse af brugeren.

Brugeren kan selv (manuelt) tillægge nye REGEX StåendeUdtryk i TextClassifyExpert.

I praksis starter brugeren med selv at klassificere (GroundTruth) på en del TekstBidder. Derefter køres TextClassify, som på ca. 5 CPU-minutter optimerer alt på en gang. TextClassify udspytter en lang række filer med alskens relevant information. En af disse indeholder søjler med renhed for hvert delmål og hvert TekstBid.

Brugeren vil typisk sortere TekstBidderne efter renhed, sammenligne med GroundTruth, og udføre følgende:

a) Undersøge FalskPositive (høj renhed uden GroundTruth). Disse skyldes enten at brugeren har overset TekstBidden (den manglende GroundTruth indsættes), eller at TextClassify tager fejl.
   Det sidste sker meget sjældent, og kan kun skyldes en stor statistisk fluktuation, eller mere sandsynligt (dog ikke set endnu), et regulært problem i algoritmen, som så kan undersøges og rettes.
b) Undersøge FalskNegative (lav renhed med GroundTruth). Disse skyldes enten at brugeren har kommet til at GroundTruth TekstBidden uden grund (den fejlagtige GroundTruth fjernes), eller at TextClassify tager fejl.
   Det sidste sker sjældent, og kan kun skyldes en stor statistisk fluktuation, eller mere sandsynligt, at en komponent i TekstBidden ikke er repræsenteret i TextClassifyExpert. Grunden til at TextClassify giver en lav renhed er tilgængelig for brugeren, som så kan bruge informationen til at identificere et eventuelt manglende StåendeUdtryk, som så indsættes i TextClassifyExpert.

c) Når a+b) er udført, genkøres TextClassify med højere kvalitet som resultat.
d) Når indlæringskurven ikke længere er stejl, afsluttes processen med en krølle på halen: Eventuelle TekstBidder med renhed omkring 50% undersøges.
Hvis disse blandede TekstBidder virkelig er svære at tildele en GroundTruth, så er optimeringen simpelthen så god som muligt. Ellers ledes der på disse TekstBidder efter et eventuelt separerende mønster i ordteksten.

Alt dette fungerer sømløst og gør det muligt for IMR, hurtigt og robust, at klassificere 60.000 TekstBiddder til cirka 90 relevante ud af 169 SDG delmål.
EkspertSystemet er dog afhængig af en EkspertBruger, som har kendskab til komponenterne i TextClassify, og som har en god forståelse for teksternes semantiske indhold, hvordan man kan forstå StåendeUdtryk og konstrurere REGEX filtre.
TextClassify kræver desuden, at brugeren selv konverterer og holder rede på .csv.txt(.xlsx) versioner. Store .xlsx filer (>10MB) er notorisk besværlige/langsomme at håndtere.

De tre EkspertBruger egenskaber kan koges ned til:
-) Behandling af StåendeUdtryk og REGEX
-) Fra-filtrering af (nogle få) af TextClassify foreslåede EkspertOrd, som åbenlyst skyldes statistiske fluktuationer.
-) Informationshåndtering

Disse 2 første egenskaber er i skrivende stund under implementering i:

PraktikantSystemet(i forventet brug 20190101)
En gennemgang af de anvendte REXEG StåendeUdtryk viser, at de alle kan koges ned til sammensætning af enkelte StamOrd via SammensatteOrd til StåendeUdtryk.

For at forstå mønstrene og semantikken i de StåendeUdtryk foretog vi en detaljeret analyse af normal engelsk "Text Corpora" mod menneskerettighedsteksterne fra FN og de StåendeUdtryk brugt som EkspertOrd. Det viser sig at de hyppigt optrædende StåendeUdtryk i FN tekster i meget stor grad består af indholdsløst formalia eller slang. Selvom den statistiske analyse giver mange resultater af akademisk interesse, er sammenligningen med Text Corpora nytteløst til vores praktiske formål.

En videre analyse af de brugbare StåendeUdtryk viser, at de altid er opbygget via simple 2->1 sammensætninger af StamOrd til SammensatteOrd som igen er 2->1 sammensatte til nye SammensatteOrd...
PraktikantSystemet kan undersøge en stor del af alle mulige sammensætninger. Når et SammensatOrd øger antallet af MaskinLærteKoblinger signifikant, udvides TextClassifyDictionary med det nye SammensatteOrd, som så igen kan sammensættes.

For at få antallet af kombinationer ned fra cirka K(Antal aktive StamOrd=20000, 2)=200 millioner til i størrelsesorden 1 million og dermed et acceptabelt CPU-brug kræves tre ting:

a) Man undgår at undersøge hyppigt forekommende ord uden semantisk mening, som for eksempel de fleste biord og verber uden reelt indhold (er, kan, vil, må, etc.). Disse ord flages i TextClassifyDictionary så de ikke spilder CPU-tiden.
b) Evalueringen af potentielle EkspertOrd foregår på alle delmål samlet, således at statistikken er størst mulig.
c) Der sker en præ-evaluering af de mulige sammensætninger uden brug af MahjongAlgoritmen. Udfra hyppigheden af to StamOrd (med/uden GroundTruth), hyppigheden af det SammensatteOrd (med/uden GroundTruth), og alle de oprindelige renheder fra MahjongAlgoritmen, beregnes en naiv Bayesisk forventningsværdi for hyppigheden af det SammensatteOrd (med og uden GroundTruth) under forudsætning af, at sammensætningen ikke er korreleret med SDG-koblingen. Forskellen mellem det observerede og forventede antal SammensatteOrd med GroundTruth udgør et effektivt præ-filter, der bruges til at nedbringe antallet af kombinationer der skal afprøves med hele MahjongAlgoritmen.

For at mimimere antallet af "falske" EkspertOrd og StåendeUdtryk, uden brug af viden fra en EkspertBruger, kræves den bedst mulige præcision på både evalueringen og dens statistiske usikkerhed. Kombinationen af hyppigt forekommende ord giver ofte store udslag i antallet af ekstra MaskinLærteKoblinger, men har også en tilsvarende større statistisk usikkerhed.

Fordi udfaldsrummet er meget lille: 64[0;3,0;3,0;3] og MahjongAlgoritmen er utrolig hurtig, er det muligt at anvende avanceret statistisk maskineri til at bestemme, om en kandidat til EkspertOrd bør anvendes eller ej.
TekstBidder har totalt 4 udfaldsværdier med et udfaldsrum på 2(GroundTruth) * 64(VSO). Efter fjernelse af et EkspertOrd får de potentielt andre værdier, men 3 af disse værdier er uforandrede, og værdien af det sidste udfald er begrænset af det oprindelige Niveau[0-3] af TekstBidden og Niveauet af EkspertOrdet[0-3]:
Dette giver 4-7 muligheder for det kombinerede udfald, før og efter fjernelsen. Det totale kombinerede udfaldsrum før og efter fjernelsen af et EkspertOrd er maksimalt 2*16*7=224. MahjongAlgoritmen udføres så på alle (maksimalt 224) udfaldsmuligheder, hvor man fjerner en enkelt TekstBid fra hele dataanalysen. Denne LOOCV/Jackknife tilgang evaluerer resultatet på TekstBidder, uden at de selv indgår i MahjongAlgoritmen, samtidigt med at variansen af resultatet (over alle tekstbidder) giver et næsten unbiased estimat af den statistiske usikkerhed.

For at lette Praktikantens arbejde viser PraktikantSystemet både tabet af MaskinLærteKoblinger og den tilhørende signifikans for hvert EkspertOrd, når dette fjernes fra TextClassifyExpert.

Informationshåndteringen har allerede været delvist implementeret i et mindre pilotprojekt udført i 2017. Dette forventes udvidet til:

BrugerVenligSystemet(i forventet brug 2019)
TextClassify er i dag et simpelt script der indlæser .csv.txt filer, tjekker al data, laver den fulde optimering, og spytter ud en stor mængde relevante men relativt uorganiserede resultater fra analyserne.
Hvis en bruger på Praktikant niveau skal kunne anvende hele systemet uden supervision kræver det, at information vises samlet, overskueligt, anvendeligt, og forståeligt!

Dette gøres lettest grafisk, hvor brugerens fokus ledes gennem farveunderlægning, hvor farven giver typen af information, og farvestyrken giver vigtigheden af informationen. I pilotprojektet udviklede Specialisterne et display der på den ene side viser den rå TekstBid. På den anden side vises fortolkningen af TekstBidden i termer af dets EkspertOrd og disses TypeNiveauer (0-9 i tilsvarende farver) for et sæt af forskellige delmål. Denne information kan brugeren så direkte sammenligne med GroundTruth informationen og renheden fra analyserne. Et enkelt display giver på den måde et fuldstændigt billede af, alt hvad der konkret foregår, for hver enkelt TekstBid, og et enkelt SDG-mål med dets delmål.

Dette display tænkes udvidet med et display, der viser et globalt billede af alle analyserne på mange TekstBidder af gangen. Her er fokus på unormale resultater, dvs. at matricen TekstBidder*delmål farvelægges således at de interessante (FalskPositive, FalskNegative, og MåskeMåske felter) vises med stærke farver.
Det giver et grafisk overblik over hvor klassificeringen kan forbedres. Samtidigt vil brugeren kunne klikke på enkelte TekstBidder, ændre GroundTruth hvis nødvendigt, og klikke på TekstBiddens komponenter for at undersøge/ændre disses egenskaber.

Et tredje Display tænkes udført for de enkelte delmål, hvor al informationen fra dataanalysen vises samlet. Desuden vil BrugerVenligSystemet kunne sortere og filtrere informationen som brugere er vandt til fra .xlsx. Samlet vil dette give et komplet brugervenligt system, hvor al information tilgås ved at klikke rundt i farverige gennemsigtige bokse.

For yderligere at forenkle Praktikantens arbejde vil TextClassifyDisplayAnalysis organisere al filarbejdet. Dette medfører også at BrugerVenligSystemet effektivt vil kunne håndtere ca. 500MB data, hvor grænsen ved brug af .xlsx i dag er cirka 20MB.

Da BrugerVenligSystemet skrives i "Visual C# Windows Forms App (.NET Framework)" er det simpelt at implementere en BrugerVenligEmulator, som på rigtige data, skridt for skridt, tager brugeren igennem alle de vigtigste muligheder.

Resultater

Resultatet af TextClassify er umiddelbart sættet af koblinger mellem anbefalinger og SDG (del)mål. Disse kan undersøges og analyseres sammen med andre metadata via en hjemmeside [UPRSDG] hos IMR. Desuden eksporteres koblingerne til brug i andre relevante databaser. Metoden bag resultaterne er præsenteret som [METHOD]. Disse sider kan ses som et pilot projekt og starten af en process. Der fremstår mange muligheder for fremtidig udvikling, enten via hjemmesiden, eller i samarbejde med andre tovholdere.

I øjeblikket omhandler fremtidsplanerne specifikt en kvantitativ analyse af indikatorerne tilknyttet SDG målene, og SDG-koblinger til andre offentligt tilgængelige tekster. Desuden er der på verdensplan stor efterspørgsel efter ekspertsystemer til at hjælpe med processering af tekstmateriale vedrørende menneskerettigheder og SDG.

Fremtidsudsigter for TextClassify

TextClassify har vist sig at være et effektivt og relativt brugervenligt maskineri til klassificering af TekstBidder til veldefinerede mål. Maskineriet er ikke sprog-specifikt. I skrivende stund er vi ved at udvide TextClassify til både spansk og kinesisk. Den simple sproglige behandling foregår indkodet i TextClassifyDictionary, mens TextClassify udfra GroundTruth mønstrene foreslår udvidelsen med SammensatteUdtryk. På nogle sprog, som spansk, halter det med officielle frit-tilgængelige StamOrdBøger, andre som dansk har dem allerede, og kinesiske tegn fungerer direkte som StamOrd i TextClassify maskineriet.

Det er let at udvide brugen af TextClassify med koblinger til andre klasser af meninger. På disse vil optræning og evaluering foregå på præcist samme måde, og den eneste forskel er valget af StamOrd som EkspertOrd, og de SammensatteOrd.

Den simple konklusion

IMR har i samarbejde med Specialisterne udarbejdet et EkspertSystem beregnet til indlæring af koblinger mellem tekster omhandlende menneskerettigheder og FN's Verdensmål. Resultaterne viser at TextClassify virker effektivt, bedre på nogle delmål end andre. Videre udvikling gør systemet uafhængigt af menneskelige eksperter og dermed skalerbart. Slutproduktet forventes at fremstå som en farverig gennemsigtig boks, der ikke bare er brugervenlig, men også i stand til at præsentere præcist hvordan de maskinlærte renheder har fået deres værdier. Grundlaget for TextClassify's succes er især MahjongAlgoritmen, der gør den multi-dimensionale dataanalyse lynhurtig.

Den indviklede konklusion

Statistikere, fortrolige med Kalman filtre og Monte Carlo baserede analyser, vil forhåbentligvis genkende mønstret i TextClassify. Alle delprocesser opfattes som stokastiske, med tilknyttede sandsynligheder. Desuden fungerer GroundTruth som GeneratorTruth i en Monte Carlo fortolkning.

Analyse		Data		Process	*Blindgyder*
SDG Mening		Ensure access to information and protect fundamental freedoms		Bæredygtig Udvikling	
Ground Truth		16.10 98% renhed		Statistisk Mening	
VSO Modellen		End Persecution HumanRightsDefender		Bruger Venlig System	
Stående Udtryk		HumanRights HumanRightsDefender		Praktikant System	*Text Corpora*
StamOrd		*Take* Immediate Step *To* End *The* Persecution *Of* Human Rights Defend		Ekspert System	
Ord		Take immediate steps to end the persecution of human rights defenders		Mahjong Algoritmen	
Tegn				VSO Modellen	*Metadata U10|D31|H1|I50*
UPR Tekster				UPR Tekster	*Databaser*

Dette paradigme gør det muligt at overføre alle de "julelege" der gennem tiderne er blevet udviklet til Monte Carlo (Kalman-)filtre. Den grundlæggende komponent er betemmelsen af kovariansen mellem data, (slut)resultatet af analysen, og GroundTruth. Alle beregninger foretages på små udvidelser/ændringer af analysen. På den måde fremstår opdateringen som et Kalman filter, hvor kædereglen for differentiation bruges i alle led. Enkelte manuelle robustificeringer er nødvendige før processen er selvkørende. Det vigtigste er selvfølgelig begyndelsesværdier for GroundTruth. Den anden komponent er StamOrdBogen, som er nødvendig på sprog med mange grammatiske former. VSO-modellen er en konkret del af maskineriet, men kan lige så godt erstattes af en (1->2)generator af undermeninger udfra det endelige delmål. På den måde kan hele maskineriet (undtagen StamOrdsGrammatikken) ende med at være uafhængig af selve opgaven, aka. "Deep Learning".

Ligesom TextClassify ligner et Invers Monte Carlo Kalman Filter, ligner udviklingen af samarbejdet, der har resulteret i TextClassify, også selve TextClassify.
Vi startede med nogle data fra nogle fejlbehæftede databaser, oprensede disse til praktisk brug, opgav databaserne, lavede VSO-modellen, udviklede MahjongAlgoritmen, og tilslut automatiserede næsten hele processen med brugervenlighed in mente.
Alle disse beslutninger er hovedsaglig foretaget baseret på kovariansen af data med dataanalysen og vores foregående viden. Også her var enkelte robustificeringer nødvendige for at opnå en effektiv proces. I vores tilfælde opgav vi brugen af de officielle databaser, og hensatte analysen af hyppigheden af mønstre i Text Corpora til et akademisk hjørne.

Referencer:

[UPR] https://www.ohchr.org/en/hrbodies/upr/pages/uprmain.aspx
[SDG] https://www.un.org/sustainabledevelopment/sustainable-development-goals/
[UHRI] https://uhri.ohchr.org/en
[UPR-INFO] https://www.upr-info.org/en
[BOW] https://en.wikipedia.org/wiki/Bag-of-words_model
[VSO] https://en.wikipedia.org/wiki/Verb%E2%80%93subject%E2%80%93object
[REGEX] https://en.wikipedia.org/wiki/Regular_expression
[GroundTruth] https://en.wikipedia.org/wiki/Ground_truth
[IPBM] https://en.wikipedia.org/wiki/Interior-point_method
[LOOCV] https://en.wikipedia.org/wiki/Cross-validation_(statistics)#Leave-one-out_cross-validation
[UPRSDG] https://www.humanrights.dk/upr-sdg-data-explorer
[METHOD] http://upr.humanrights.dk/methodology

How Machine Learning helps doctors prevent patients from getting hospital acquired infection

Authors: J. K. Møller[A], C. Hardahl[B], M. Johansson[B], M. Sørensen[B] and G. Bergman[B]

A) Department of Clinical Microbiology, Lillebaelt Hospital, Vejle, Denmark
B) SAS Institute

WORK IN PROGRESS, DO NOT REFER TO OR CITE WITHOUT PERMISSION FROM THE AUTHORS.
THIS VERSION: JANUARY 4, 2019

Abstract
Hospital-acquired infections (HAI) are the most frequent adverse event in healthcare worldwide. 7-10% of all hospitalized patients will acquire at least one hospital-acquired infection:

Europe: 3.2M patients/year, 37 000 deaths/year
US: 1.7M patients/year, 99 000 deaths/year

Infections stands for 35% of all medical errors in the Western world according to WHO 2016. Despite of antibiotic treatment, there is a mortality rate of 15-20% for serious HAI. Our results from Nordic Hospitals shows that patients with a HA-UTI (Hospital Acquired Urinary Tract Infection) is hospitalized on average twice as long as patients that do not get a HA-UTI. (9.8 days versus 4.4 days)

Region of Southern Denmark (RSDK) has together with SAS Ins. build a tool to predict which patients have the highest probability of acquiring a HAI during hospitalization. More than 300.000 hospitalizations have been included in the cohort where data has been extracted from unstructured Electronical Medical Record notes, structured data like medicine, ICD10-codes, LAB-results and demographical data.

RSDK will use the tool to initiate a series of programs to prevent patients from getting HAI. For this article we have moved from the "simple" transparent decision tree model to a complex ML algorithm with Decision Trees and bagging to further enhance model performance. With the promising results from the Machine Learning (ML) algorithms we will hopefully soon see this part of the tool helping patients at risk.

1 Background

Healthcare associated infections (HAI) are a major burden for the healthcare system and are associated with prolonged hospital stay, increased morbidity, mortality and costs. Hospital acquired urinary tract infections (HA-UTI) accounts for about 20-30% of all HAI according to (Zarb et al., 2012) and (Burke, 2003), and with the emergence of multi-resistant urinary tract pathogens, the total burden of HA-UTI will most likely increase.

Other studies have identified several risk factors for developing HA-UTI. (Redder et al, 2016) identifies admission with community acquired urinary tract infections (CA-UTI), diseases in the genitourinary system and diseases of the nervous system as significant factors influencing the odds of developing a HA-UTI. (King et al, 2012) presents a systematic review, calculating Population Attributable Risk (PAR) for factors included in 23 papers. Factors with a high and significant PAR include urinary catheter, stroke, female sex and hypertension.

This paper is a part of a larger study where the primary aim is to create computerized algorithms for continual surveillance of the incidence of hospital acquired infections. A natural extension to that aim is to create a model predicting the risk of an admitted patient developing a CA-UTI or HA-UTI.

When the model is sufficiently accurate and robust it can be deployed and used in the daily routine at hospital care units, identifying potential risk patients, allowing the care providers to take proper measures to lower the risk. For example, risk scores may be added to the daily rounds where the care of each patient is evaluated and planned further.

A good model and a right user interface for the care providers will mean less time needed to read through all background information around a patient and decisions based on much more experience/learning than a single care provider can obtain through his/her lifetime.

2 Hypothesis

Our hypothesis is that CA-UTI and HA-UTI can be predicted at admission time of a hospitalization using a mix of structured and unstructured data from Electronical Medical Record (EMR) notes, medicine intake, ICD10-codes, LAB-results and demographical data.

2.1 Aim

The overarching aim of this paper is to show the potential of mathematical models to predict urinary tract infections at the admission time of a hospitalization (CA-UTI and HA-UTI). For the caretaking staff to prevent patients from acquiring UTIs, they need to know why our model return a high probability for a patient to get infected. Therefore, the main goal is to

1. Develop a model that predict the probability of acquiring any UTI (CA-UTI or HA-UTI) at admission using a transparent and interpretable model.

When the caretaker staff has gotten used to the model, and start believing it is correct, we want another model, where we sacrifice interpretability for accuracy. The secondary goal is to

2. Develop a black-box machine learning model and compare the results to the more interpretable model.

3 Data source

The data foundation consists of multiple clinical systems with both structured and unstructured data variables.

Clinical systems include:

- Cosmic Electronical Medical Record (EMR)
- Cosmic Patient Administrativ System (PAS)
- Cosmic medicin module
- Microbiology results (MADS-databse)

The database consists of both structured and unstructured data. Data from intensive care unit are not included. The clinical systems include information about the patient and information about the admission including data on diagnoses, procedures, administered drugs, prescribed drugs, microbiology results, other lab results.
Data from the EMR is unstructured and needs to be converted from text to usable structured variables that work as risk factors in the predictive model. Examples of risk factors is fever registrations (> 38 degrees) and pain when urinating. These types of risk factors can be described in various ways and with different words, synonyms and medical slang and this must be considered and validated when creating structured variables from unstructured data.
The sampled data set described in the next section is based on an automated HAI/CAI UTI screening algorithm that screens all admitted patients in the Region of Southern Denmark. The sampled data set uses the screening algorithms results and extracts admissions with CA-UTI or HA-UTI to form our case data to be used in the predictive model (target variable).

3.1 Sampled data

The sampled cohort consist of all patients discharged between 01 Jan 2017 and 31 May 2018 in Region Southern Denmark, consisting of 4 hospitals: Sygehus Lillebælt, Odense Universitetshospital, Sygehus Sønderjylland and Sydvestjyst Sygehus. The sample consist of 301,932 observations, where each correspond to a unique patient hospitalization.

In case there are several episodes that fulfil the criteria of either HA-UTI or CA-UTI during a single care episode, a window of 14 days after the first positive event will be applied where no new events can be registered. If the criteria are met once again after those 14 days, a new event of UTI can be registered, similar to the Repeat Infection Timeframe (RIT) of the CDC criteria (PSCM, 2017).

3.2 Variables

In this section we list and explain the variables used as input- and response variable in the models. The variables are all known at admission time of the patient.

3.2.1 Target variable definition

The definition of a HA-UTI is a urinary tract infection discovered in a hospitalized patient, which at the time of discovery, has been hospitalized more than 48 hours. A CA-UTI is a urinary tract

infection discovered in a hospitalized patient which at the time of discovery has been hospitalized less than 48 hours.

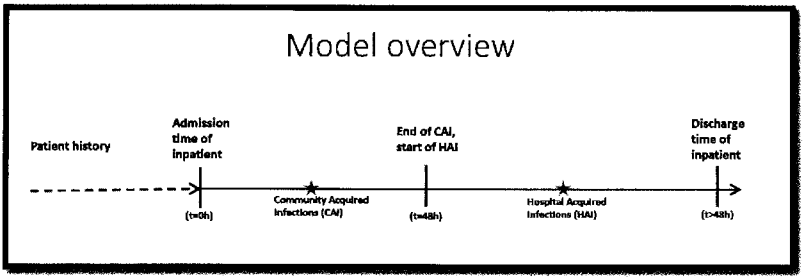


Figure 1: Illustration of the time lines of the two predictive models; the Entry Model and the HA-UTI model.

The target variable is UTI-Entry, which is created from HA-UTI and CA-UTI. If the patient acquires any of the 2 UTIs during hospitalization, then UTI-Entry equals 1 else it equals 0. We have made a model entirely for HA-UTI, which can be discussed another time.

3.2.2 Descriptive statistics

We have selected input variables of interest based on Doctors and caretakers experience with UTIs combined with studies in the area (Redder et al, 2016) and (King et al, 2012).

Variables	**Total**
Unique Admission ID', N	301,932
Target Entry model (CAI eller HAI), %	6.3%
Male, %	46.2%
Age, mean (SD), N	51.6 (27.4)
Admission Type, Emergency, %	70.8%
Admission Hospital	OUH: 40.7% SVS: 15.7% SLB: 25.8% SHS: 17.8%
Readmission, %	8.6%
ICD10, diabetes, at admission, %	3.2%
ICD10, urinary retention, at admission, %	0.7%
ICD10, kidney (DN18) , at admission, %	1.4%
ICD10, stroke, at admission, %	1.9%
ICD10, hypertension, at admission, %	5.5%
ICD10, unconsciousness, at admission, %	1.1%

ICD10, KOL, at admission, %	3.1%
ICD10, atrial fibrillation, at admission, %	0.4%
ICD10, unresolved spinal, at admission, %	0.0%
ICD10, CVK start, at admission, %	1.1%
ICD10, CVK change, at admission, %	0.0%
ICD10, CVK end, at admission, %	0.0%
ICD10, bladder scan, at admission, %	2.4%
ICD10, neuro disease, at admission, %	2.7%
Previous_HA_UTI, at admission, %	2.4%
Previous_CA_UTI, at admission, %	8.0%
Referral_diagnosis	A lot of levels
Admitted_org_id	A lot of levels

The ICD10 variables are diagnosis codes.

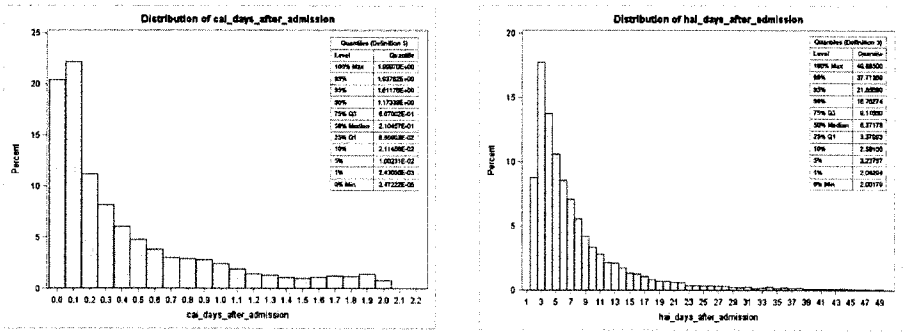


Figure 2: The distribution of time from admission to acquiring a Community acquired or Hospital acquired infection respectively

Most infections happens right after admission, as depicted in figure 2. This is an interesting finding, as that makes predicting the probability of getting an UTI at admission even more relevant.

4 Method

Given the nature of our data, we make use of supervised machine learning algorithms to model the response variable. We split the data into 2 datasets: Modelling data and out-of-time validation data. The modeling data consists of hospitalizations where the discharge date is between 01 Jan 2017 00:00 and 01 May 2018 00:00 and the out-of-time consists of discharge dates between 01 May 2018 00:00 and 01 Jun 2018 00:00.

To achieve the goal of creating an easily interpretable model, we have decided to model the data using a Decision Tree. This model is used to give an indication of what parameters has a big effect on acquiring a UTI and is used as a benchmark for the more advanced machine learning methods.

When we build the black-box model, instead of building an entirely new model, we have used a similar technique as Breiman's Bagging approach for classification Trees (Breiman, 1996), to enhance the performance of our initial model. (Bühlmann & Yu, 2002) has later proven the technique lowers the variance of the model, and hence makes the model more robust when scoring new observations. We'll take advantage of (Bühlmann & Yu, 2002) approach, where subsamples are used instead of the first implementation with bootstrap.

In the coming section we will discuss how we treat patients still admitted after the end of the data collection period. Then we investigate the decision tree approach and the more advanced bagging method.

4.1 Censored observations

Patients who are still hospitalized, and who have not been infected with a HA-UTI or CA-UTI at the end of the sample period is right censored. Due to the nature of the machine Learning algorithms used, we do not have any good way to model these, and therefore we have chosen to remove them from the sample. This type of observation represents less than 0.1% of the data, so the effect of removing them is thought to have minimum impact on the models.

4.2 Modelling

The SAS Enterprise Miner implementation of decision trees finds multi-way splits based on nominal, ordinal, and interval inputs. The algorithm is based on the Chi-square automatic interaction detection (CHAID) method in (Breiman et. al., 1984). The splitting criteria is based on p-values from the Chi-square distribution (Binary targets).

We have chosen a simple decision tree with a binary split and a maximum depth of 6, as we want the tree to be easily interpretable for non-statistical thinkers. We have tried using trees with three splits, more depths and so on, and the results has been near the same.

We divide the modelling data into two: 60% for training and 40% for validation. This is to lower the chances of overfitting our model. In turn we have not used cross-validation.

To increase model accuracy, we sacrifice some model interpretability, and implement a subagging-method as described in (Bühlmann & Yu, 2002). In the end we evaluate the models' performance on the out-of-time dataset and investigate if the loss of interpretability has given extra performance.

The algorithm we use:
Step 1: For $k = 1, \ldots, 10$ do:

1. Generate a random sample from the modeldata, $(X_1^k, Y_1^k), \ldots, (X_m^k, Y_m^k)$, using simple random sampling without replacement m times for $m = 70\%$ of the modelling data.
2. A decision tree is build using the training dataset with the same specifications as in the baseline model, but now with a 10-fold cross-validation. The output estimatior is given as $\hat{y}_k$.

Step 2

1. The subagging model output is the average of all the model outputs from step 1. Given as

$$\frac{1}{k}\sum_{i=1}^{k}\hat{y}_i$$

Normally you would choose $k = 100$ or even higher, but because of computation time and the limited use of the model further on, we have chosen to use only 10 models. This is subject to change.

We have decided to evaluate the model comparison using Area Under Curve of the Receiver Operating Curve, average square error and the cumulative lift. In the next chapter we will investigate our findings and evaluate model performance for each of the 2 models.

5 Results

From the baseline model we have found certain attributes show for patients with a high probability of acquiring an UTI:

- Elderly patients (Patients above the age of 75 are 4 times more likely to acquire an UTI)
- Patients with previous UTI infections are more likely to get infected again
- The probability depends on which department the patient is hospitalized in (ie. patients in certain departments are more likely to get an UTI)
- Patients who are hospitalized acute
- Women are more likely to get infected than men
- Patients who previously have had an indwelling catheter

We have made a table with the model measurements for the baseline and ensemble model for both training and out-of-time validation.

Table with performance measurements

Model	AUC	ASE	Lift
Baseline, Training	0.819	0.052	4.03
Baseline, oot-val	0.818	0.055	3.98
Bagging, Training	0.830	0.052	4.15
Bagging, oot-val	0.829	0.054	4.09

The advantage of using the ensemble model compared to the baseline model has not given much performance. We sacrifice a lot of interpretability for a .01 increase in AUC, the same ASE and a

slight increase in lift. This is most likely due to the ensemble process needing more models to really improve. None of the models are overfitting the training data and perform quite well on the out-of-time validation dataset.

6 Discussion

Some of the found risk factors was well known beforehand, i.e. women have a more exposed urethra than men, and are therefore more exposed to bacteria. The advantages of these models will be seen when they are put into production, and caretakers start using them in their daily work. For example, Odense Universitetshospital has hired a specialist to do all catheter placements, to minimize the risk of doing it wrong, which can lower the infections due to indwelling catheters.

The model performance did not improve much when we introduced a more advanced modelling technique. We still need to test what effect the bagging method have had on the model's variance and decide if it is worth continuing with that kind of model.

For future work, we will experiment with models that are more state-of-the-art and even more advanced like neural networks and gradient boosting. We have a strong belief that these models will outperform the baseline model and make hospitals more secure for all of us.

7 References

(Breiman et. al., 1984), Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. Classification and Regression Trees. Wadsworth International Group, Belmont. (1984)

(Breiman, 1996), Breiman, L.: Bagging predictors. Machine Learning, 24, 123–140 (1996)

(Burke, 2003), Burke, J. P. Infection Control — A Problem for Patient Safety. N. Engl. J. Med. 348, 651–656 (2003).

(Bühlmann & Yu, 2002), Bühlmann, P., Yu, B.: Analyzing bagging. Annals of Statistics 30, 927–961 (2002).

(King et al, 2012), C. King, L. Garcia Alvarez, A. Holmes, L. Moore, T. Galletly, P. Aylin. Analysing risk factors for urinary tract infection based on automated monitoring of hospital-acquired infection. Journal of Hospital Infection, 82: 219-226 (2012).

(PSCM, 2017), Centers for Disease Control and Prevention. National Healthcare Safety Network (NHSN) Patient Safety Component Manual. (2017).

(Redder et al, 2016), J.D. Redder, R.A. Leth, J.K. Møller. Analysing risk factors for urinary tract infection based on automated monitoring of hospital-acquired infection. Journal of Hospital Infection, 92:297-400 (2016).

(Zarb et al., 2012), Zarb, P. et al. point prevalence survey of healthcare-associated infections and antimicrobial use in European acute care hospitals. Eurosurveillance 17, (2012).

PCA-based exploratory data structure comparisons – or *A Story of Happiness Rankings*

Anne Helby Petersen[1]*, Bo Markussen[2] & Karl Bang Christensen[1]

[1]Department of Public Health, University of Copenhagen; [2]Department of Mathematical Sciences, University of Copenhagen

Abstract: Many datasets are somehow divided into distinct subsets, e.g. due to multi-center sampling, survey experiments or multiple measurement instruments being applied. However, we often still wish to analyze the data jointly, and therefore, we should consider whether the data subsets are sufficiently similar in structure for this endeavor to be meaningful. We propose to address this question of data structure similarity by use of three new diagnostic plots, all based on principal component analysis. We apply these exploratory methods to data from the European Social Survey measuring psychological well-being and discuss whether or not this concept is comparable across countries, as is often done e.g. in the UN Happiness Report, where Denmark is always in the top. Are we really happier than everybody else, or could it be that the UN happiness definition is just designed to measure our kind of happiness?

1 Introduction

Data comparability is a recurring topic in applied statistics. Often, data are collected in such a way that it is essentially divided into several subsets whose comparability needs to be assessed empirically. This happens for instance when data are collected across several centers (or countries) or when different versions of an instrument, the mode of administration of a survey, or the order of survey questionnaire items are applied. An example of where the data structure comparability issue must be considered is when conducting international rankings of countries according to e.g. educational quality (e.g. the PISA project) and citizen happiness (e.g. the UN *World Happiness Report* project). From a methodological point of view, such international rankings are problematic, as they rely on the fundamental assumption that the measured concepts are inherently the same across countries.

The comparability issue in these examples can be summarized as follows: Assume that we have two datasets with the same variables, but different observations, often represented as a single dataset with a subset-inducing variable, and that we wish to compare them without specifying a model, or even a variable of interest. The central question is then whether the two datasets can readily be combined for the purpose of later data analysis, or if the subset-inducing variable implies heterogeneity that must be dealt with in later statistical modelling. Sophisticated methods for addressing this question are available when we are willing to assume a statistical model. However, this places the effort of assessing data comparability very late in the data analysis process and it makes the comparability assessment ad-hoc, as it essentially relies on the appropriateness of the modeling choices, which again depends on the structure of the data. We propose three diagnostic plots that can be used for

*Corresponding author: ahpe@sund.ku.dk

data structure comparisons without having to specify a statistical model. These methods use principal component decomposition of the empirical covariance matrix in the two subsets to create intuitive visualizations of data structure differences and we refer to them collectively as *Principal Component Analysis-based Data Structure Comparisons* (PCADSC). Note that the use of PCA implies that we can only compare datasets with numerical variables and that we focus on the linear aspect of the data structures.

We will showcase how the plots can help detect and understand data structure differences by applying them to data on psychological well-being from three European countries that may or may not be comparable. Our starting point is Denmark which has repeatedly been awarded with the title of "happiest country in the world" by the *World Happiness Report*, most recently in 2016 [3], and we wish to investigate if this title is really meaningful at all. In the rankings of happiness, not much work has yet been devoted to evaluating the assumption of international comparability, though Veenhoven presents a theoretically thorough, but empirically simplistic, summary of possible reasons for lack of comparability [9], while Lolle & Andersen show highly potent translation issues for the term *happiness* [6]. We will address the question by focusing on a single aspect of happiness, namely psychological well-being and we will compare Denmark with Bulgaria and Sweden. Though both Denmark and Bulgaria are European countries and thus neither very far apart geographically nor culturally, these two countries have previously been highlighted to be very different in terms of what defines happiness [4]. Moreover, intra-European regional differences in the relationship between social capital and happiness have also been demonstrated; when comparing Northern European countries to other European countries, a much less pronounced relationship between the two concepts is generally found [8]. In particular, interpersonal relations play a less important role in Denmark, compared to Bulgaria. Therefore, a successful method for data comparisons should be able to detect these differences by looking at data on psychological well-being from these two countries. We also compare the Danish data with Swedish data in order to investigate if the PCADSC tools actually do hold discriminatory power. Denmark and Sweden are often deemed very similar in terms of culture and history. Therefore, we expect fundamental concepts, such as psychological well-being, to be similar across these two countries.

2 Methods

We will use the three diagnostic PCADSC plots to compare the covariance matrices of the Danish and Bulgarian data, and the Danish and Swedish data, respectively. Note that if all variables in the subsets are jointly normal with known means, covariance matrices are sufficient statistics describing the joint distribution of all the variables, but even without the normality assumption, pairwise correlations and marginal variable variances are still interesting quantities describing linear interrelations between variables. This makes the empirical covariance matrix a reasonable place to start looking for differences in data structures. Computing empirical covariance matrices for the two subsets and comparing them entry-by-entry becomes increasingly difficult as the number of variables increases and for this reason we decompose and recompose the covariance matrices using *principal component analysis* (PCA) to get an overview of differences between the two subsets.

We propose the following three diagnostic plots:

The CE plot: The CE (cumulative eigenvalue) plot can be used to illustrate differences in eigenvalues, that is, in the relative importance of the directions identified by the PCA. The CE plot is accompanied by two permutation-tests (using the *Kolmogorov-Smirnov* and the *Cramér-von Mises* test statistics, respectively), testing the hypothesis of no difference in variance components. This plot displays a cummulative eigenvalue curve together with a 95% confidence region constructed under the null-hypothesis of no difference in eigenvalues. A CE curve that

falls within the confidence region thus suggests similar eigenvalues.

The angle plot: The angle plot compares both loadings and eigenvalues at once and it can be used to understand the information loss if the data structure of one dataset is superimposed on the other, thereby revealing which principal components (i.e. loading and variance component pairs) that are most similar and most different across the two datasets. The plot displays an angle for each pair of loadings from the two datasets under comparison, and the size of the angle is proportional to the correlation between the relevant loadings. The lengths of the lines represent the sizes of the eigenvalues, which means that angles with longer lines correspond to components that explain more variance. The angles are accompanied by shaded regions displaying the distribution of the angles under the null-hypothesis of no difference in data structures. Angles that fall within the shaded regions thus support a conclusion of similar data structures.

The chroma plot: The chroma plot is primarily an illustration of the loading patterns and it targets the question of how the roles of the original variables are different between the two datasets, thus leading the data structure comparison question back to its original, empirical context. Each component of interest is illustrated with a horizontal bar, divided into blocks. Each block represents a variable and the size of the block represents its weight in the loading vector of the principal component. Thus, similar overall visual impressions will imply similar loading structures.

Further details and a full mathematical specification of the PCADSC plots are available in Appendix A. We suggest that the plots are applied in the order in which they were presented, thereby moving from overall assesments of similarity to explanations about the nature of potential data structure differences. This workflow is summarized in Figure 1. The plots are all implemented in the `R` package `PCADSC` [7], which is available on CRAN, and can be tried out very easily using only a few lines of code:

```
library(PCADSC)
results <- PCADSC(data, splitBy = "category")
CEPlot(results) #make CE plot
anglePlot(results) #make angle plot
chromaPlot(results) #make chroma plot
```

Here, `data` is a dataset where the two subsets to be compared are indicated by a variable named `category`.

3 Comparing psychological well-being across countries

We use data from the 2012 version of the European Social Survey (ESS) project to investigate inter-country differences in psychological well-being and happiness.

3.1 Data

The ESS 2012 data contain a total of 626 variables collected from 54673 citizens of 29 countries. Here, we will only work with a subset of 35 questionnaire items that are all related to psychological well-being. These 35 items can be divided into 6 distinct scales, namely *Evaluative well-being*, *Emotional well-being*, *Functioning*, *Vitality*, *Community well-being*, and *Supportive relationships*. More details on these scales can be found in [4]. We represent each of the scales by a single variable, calculated as the average score on the items related to

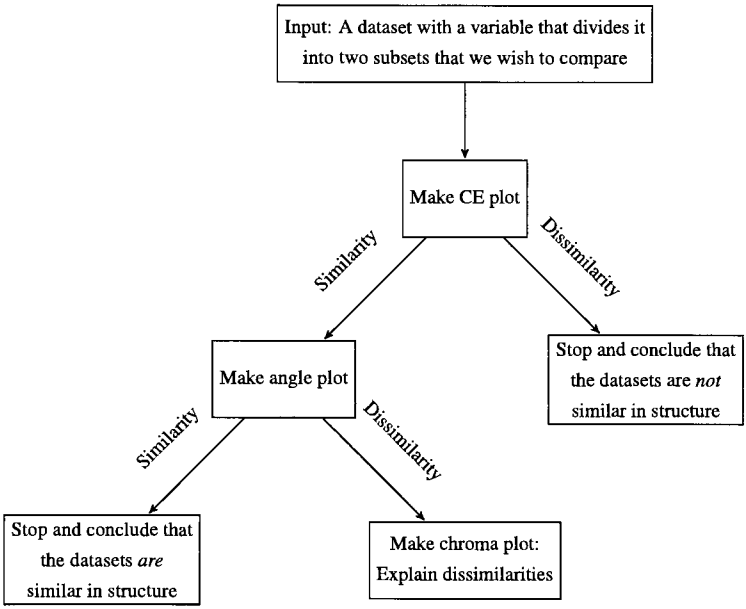


Figure 1: The PCADSC workflow, starting with two datasets and then successively applying the three visual tools for obtaining a deeper and deeper understanding of similarities and dissimilarities in the data structures.

	Denmark	Bulgaria	Sweden
Evaluative well-being	8.75 (8.00, 9.50)	5.00 (3.50, 7.00)	8.00 (7.00, 9.00)
Emotional well-being	8.33 (7.22, 8.89)	6.67 (5.00, 7.78)	7.78 (6.67, 8.89)
Functioning	7.57 (6.93, 8.21)	6.68 (5.50, 7.68)	7.04 (6.39, 7.68)
Vitality	7.50 (6.67, 8.33)	7.50 (5.83, 8.33)	7.50 (6.67, 9.17)
Community well-being	6.77 (5.83 ,7.57)	4.67 (3.70, 5.70)	6.57 (5.66, 7.37)
Supportive relationships	8.25 (7.42, 8.92)	7.25 (6.17, 8.08)	8.25 (7.42, 8.75)

Table 1: Median values of each of the six dimensions of psychological well-being, stratified by country. 25 and 75 percentiles are listed in parentheses. The scales are constructed such that they all run from 0-10.

that variable and scaled to take values between 0 and 10. For simplicity, we use only complete cases for this construction and thus exclude all participants that did not answer all the 35 questionnaire items used below. This corresponds to approximately 9% of the observations in the Danish sample, 20% in the Bulgarian sample and only 6% in the Swedish sample. All in all, we have $n_{DK} = 1498$ complete case observations in the Danish sample, $n_{BG} = 1798$ observations in the Bulgarian sample and $n_{SE} = 1736$ Swedish observations. Table 1 summarizes the marginal distributions of the six dimensions of psychological well-being, stratified by country.

3.2 Comparing Denmark and Bulgaria

Figure 2 presents the CE plot and the angle plot obtained from comparing the Danish and Bulgarian psychological well-being scales. The CE plot show a remarkable degree of lacking comparability: The cumulative differences in the eigenvalues by far exceed what could come about randomly if there really were no difference in the data structures. This is also confirmed by the Kolmogorov-Smirnov and the Cramér-von Mises tests, which both result

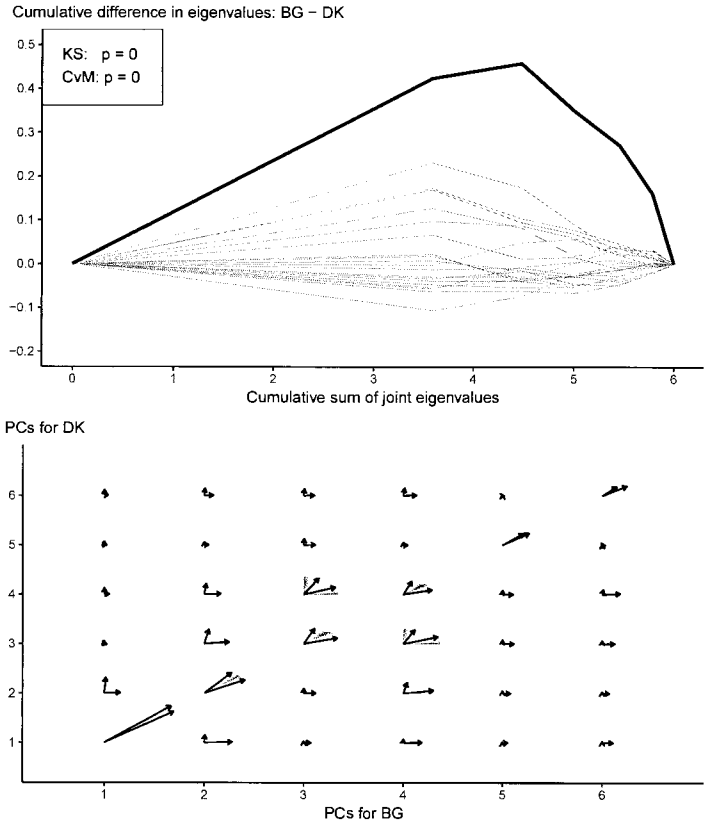


Figure 2: CE plots (top panel) and angle plots (bottom panel) comparing Danish and Bulgarian data on psychological well-being. The CE plot is annotated with the p-values of the Kolmogorov-Smirnov and the Cramér-von Mises tests of the hypothesis of no difference in data structures. In the angle plot, the upper arrows show the principal components of the Bulgarian dataset decomposed in the coordinate system of the principal components of the Danish dataset, while the lower arrows illustrate the reverse.

in p-values that are virtually zero. The CE curve generally lies above zero, suggesting that the Bulgarian dataset has larger eigenvalues for all but the last component, thereby explaining more variance at lower dimensionalities.

Moving on to the angle plot, we find that the differences are primarily to be found in the second, third, and fourth principal components (PCs). The upper arrows visualize the decomposition of the principal components for the Bulgarian dataset in the coordinate system of the Danish dataset. We see that PC2 also loads on PC3, that PC3 also loads on PC4, and that PC4 also loads on PC2 and PC3. The lower arrows visualize the decomposition of the principal components for the Danish dataset in the coordinate system of the Bulgarian dataset. Here, we see that PC2 also loads on PC4, that PC3 also loads on PC2 and PC4, and that PC4 also loads on PC3. Thus, if we wish to understand why differences in the data structures occur, an inspection of the loadings of components 2, 3 and 4 might be informative.

The chroma plot in Figure 3 allows us to look closer into these components. Here, we find that the relative importance of the *Community well-being* and *Supportive relationships* scales is much larger in the Bulgarian sample than in the Danish. In the Danish data, on the other hand, we find that *Vitality* and *Emotional well-being*

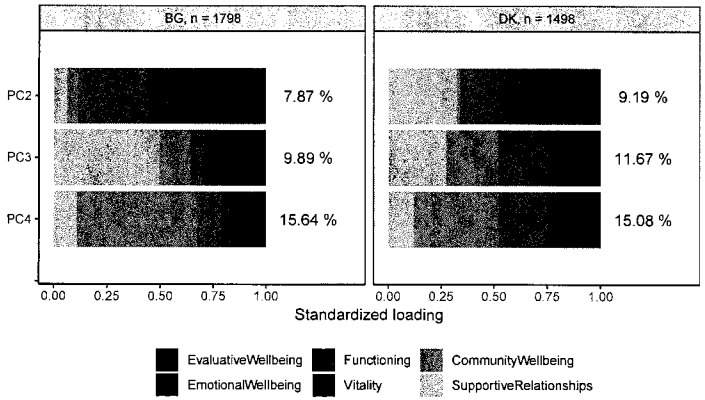


Figure 3: Chroma plot for comparing Danish and Bulgarian data on psychological well-being. A chroma plot comparing the 2nd, 3rd, and 4th principal components of the Bulgarian- and Danish psychological well-being data. The component-bars are annotated with their relative variance contributions.

seem to play bigger roles, as they appear with larger loadings in more high-ranking components in this sample, relative to the Bulgarian.

All in all, we find that psychological well-being does *not* seem to be the same concept in Bulgaria and Denmark. The two countries disagree both in how many dimensions are needed to capture the most important parts of the concept (as illustrated by the differences in eigenvalues) and in how these dimensions are then weighted among the 6 scales (as illustrated by the angle- and chroma plots). In Bulgaria, interpersonal features seem to be more informative of psychological well-being, whereas in Denmark, individual characteristics play a relatively larger role, corresponding well with the previous findings mentioned above. Thus, the datasets are fundamentally different and we should therefore be wary about combining them in a joint analysis, which was also the conclusion of the ESS authors, though based on country-level aggregated statistics [4]. Moreover, this also implies that two countries cannot be ranked in terms of which country is "the most happy", at least not by referring to psychological well-being dimensions such as those encountered here.

3.3 Comparing Denmark and Sweden

We now turn to the comparison of Denmark and Sweden in terms of psychological well-being. Figure 4 shows the CE- and angle plots for these two countries. In the CE plot, we now find the cumulative eigenvalue curve to be just within the acceptance region of the null-hypothesis. This is also reflected by the two tests, which now produce p-values of $p_{\text{KS}} = 0.17$ and $p_{\text{CvM}} = 0.11$, respectively. Thus, it is not unreasonable to assume equal eigenvalues in the two datasets.

The angle plot in Figure 4 shows that the two datasets agree very strongly about the relative importance of the six scales in the six PCs, as almost all off-diagonal arrows are practically non-existent. This implies that if one already has the information held in the first PC from the Danish data, this information is in itself mostly sufficient to describe the first PC of the Swedish data.

Looking at the chroma plot in Figure 5, the same tale is told once again: Here, we find remarkably similar loading patterns in the first three components (which are responsible for almost 80% of the variance in both datasets), and slight, but increasing, differences in the remaining three components. We therefore conclude that

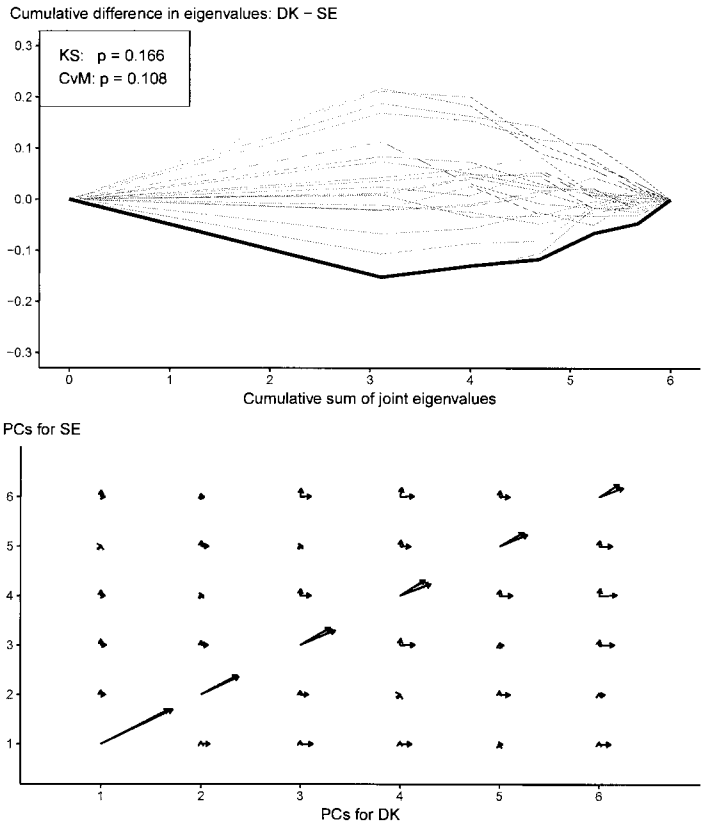


Figure 4: A CE plot (top panel) and an angle plot (bottom panel) comparing Danish and Swedish data on psychological well-being. The upper arrows show the principal components of the Danish dataset decomposed in the coordinate system of the principal components of the Swedish dataset, and the lower arrows illustrate the reverse.

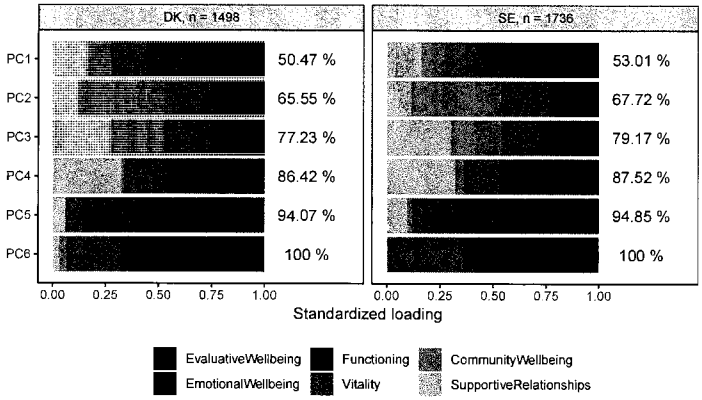


Figure 5: Chroma plot comparing the loading patterns of the Danish and the Swedish subsamples. For each component, the bar is annotated with its cumulative variance contribution, that is, how much variance can be explained by having information of this and the preceding components.

any differences in the data structures of the Danish and the Swedish samples are related to the least important dimensions of the datasets and that these dimensions are only responsible for less than 25% of the variance in both datasets. In particular, this means that we can combine and compare the Danish and Swedish datasets in a meaningful way and for instance conclude using Table 1 that in general, Danes seem to be somewhat more happy than Swedes, and in particular that the least happy people in Denmark (represented by the 1st quartiles) are generally happier than the least happy people in Sweden. A more thorough, statistical investigation could now be put to work on answering *why* this seems to be the case.

4 Conclusion

We have presented three new tools for visual, exploratory data structure comparisons that utilizes covariance matrix decompositions obtained by use of principal component analysis, and used them to investigate whether comparisons of psychological well-being scores are meaningful across Denmark, Bulgaria and Sweden. We found that while Denmark and Sweden seem to have similar concepts of psychological well-being, Denmark and Bulgaria do not. This means that international rankings of psychological well-being are not justified. And this may further imply that rankings according to more broad concepts that rely on psychological well-being, such as *happiness*, may not be meaningful.

Appendix A: Mathematical details of PCADSC

The tools presented here are all based on comparing the PCA results across two different datasets that contain the same variables. In our approach, we standardize the variables before conduction the PCA, which is equivalent to working with the empirical correlation matrices rather than the empirical covariance matrices. We also assume that all considered variables have a numerical interpretation, for example by being continuous or ordinally categorical. Moreover, we use the following notation: $\mathbf{X} \in \mathbb{R}^{n_x \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n_y \times d}$ are datasets containing the same number of variables, d, but possibly different numbers of observations, n_x and n_y. We use X_j to refer to the

jth variables of $\mathbf{X}$ and x_{ij} to refer to the ith observation within that variable, while $x_{i\cdot}$ is the full ith observation row. Note that $X_j \in \mathbb{R}^{n_x}$, $x_{ij} \in \mathbb{R}$, and $x_{i\cdot} \in \mathbb{R}^d$. We let $\bar{X} = (\bar{X}_1, ..., \bar{X}_d)^T = \frac{1}{n_x}\sum_{i=1} x_{i\cdot}$ denote the variable averages and use $S_x = \frac{1}{n_x - 1}\sum_{i=i}^{n}(x_{i\cdot} - \bar{X})(x_{i\cdot} - \bar{X})^T \in \mathbb{R}^{d\times d}$ to denote the empirical covariance matrix of $\mathbf{X}$. We let

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \in \mathbb{R}^{(n_x+n_y)\times d}$$

denote the combined dataset.

For each of the three datasets, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, we complete the following steps (here described for $\mathbf{X}$ only):

1. Standardize each of the variables to have mean zero and unit standard deviation. Let $\tilde{\mathbf{X}}_j \in \mathbb{R}^{n_x\times d}$ be the standardized dataset.

2. Form the principal component analyses

$$S_x = \frac{1}{n_x - 1}\sum_{i=1}^{n_x} \tilde{x}_{i\cdot}\tilde{x}_{i\cdot}^{\top} = \sum_{j=1}^{d} \lambda_j^x \eta_j^x (\eta_j^x)^{\top}$$

   thereby obtaining loadings η_j^x and eigenvalues λ_j^x for $j = 1, \ldots, d$.

The hereby obtained PCA decompositions of the correlation matrices can then be compared. The standardization implies that the diagonal elements of S_x, S_y, and S_z all equal 1, and thus also that

$$\sum_{j=1}^{d} \lambda_j^x = \sum_{j=1}^{d} \lambda_j^y = \sum_{j=1}^{d} \lambda_j^z = d.$$

This identity will simplify some expressions below.

The cumulative eigenvalue plot

The cumulative eigenvalue (CE) plot compares the eigenvalues of the correlation matrices. These eigenvalues represent how much information is withheld in each component in terms of explained variance. Thus, by comparing cumulative sums of eigenvalues, it is possible to obtain a detailed picture of how the two datasets differ in terms of what components are the most informative. In order to investigate whether the same proportion of the total variation can be described by the same number of principal components in the two datasets, we plot a piecewise linear curve connecting the points

$$(0,0),\ (\lambda_1^z, \lambda_1^x - \lambda_1^y),\ (\lambda_1^z + \lambda_2^z, \lambda_1^x + \lambda_2^x - \lambda_1^y - \lambda_2^y),\ \ldots,$$
$$\left(\sum_{j=1}^{d} \lambda_j^z, \sum_{j=1}^{d} \lambda_j^x - \sum_{j=1}^{d} \lambda_j^y\right)$$

Due to the standardization, the last point will always be equal to $(d, 0)$. Thus, this curve will begin and end at the x-axis. And the larger excursions it makes away from the first axis, the less alike the cumulative eigenvalue sums for the two datasets are. Moreover, a positive cumulative difference implies that dataset $\mathbf{X}$ holds more information in the first components than dataset $\mathbf{Y}$ does. By using the cumulative eigenvalues of $\mathbf{Z}$ as the first coordinate, we obtain a visualization where a component takes up an amount of horizontal space that corresponds to the variance explained by the component (that is, its eigenvalue), thereby letting the more influential components drive the visual impression.

In order to test whether these cumulative differences are statistical artefacts or if they represent something real, we have implemented both the *Kolmogorov-Smirnov* and the *Cramér-von Mises* test statistics, which are given by

$$\mathrm{KS} = \max_{k=1,\dots,d} \left| \sum_{j=1}^{k} \lambda_j^x - \sum_{j=1}^{k} \lambda_j^y \right|, \qquad \mathrm{CvM} = \sum_{k=1}^{d-1} \frac{\lambda_k^z + \lambda_{k+1}^z}{2} \left(\sum_{j=1}^{k} \lambda_j^x - \sum_{j=1}^{k} \lambda_j^y \right)^2.$$

We conduct the tests as *permutation tests*, that is, by randomly reallocating the combined and separately standardized datasets into two new datasets of n_x and n_y observations, respectively, and then redoing the CE plot steps and recalculating the test statistics. This should be done a large (e.g. 10000) number of times. Then, a p-value is obtained by computing the proportion of reallocated datasets that lead to test statistics at least as large as the one we found for the original datasets.

The permutation test results are also used to visualize the uncertainty of the CE curve in the plots. In the CE plots shown in the following section, we plot the observed curve together with 20 of the resampled curves, as well as a shaded region visualizing pointwise 95 % coverage intervals. If the observed curve is very different from the resampled curves or if it is substantially outside the shaded region, then this also indicates differences between the two datasets.

The angle plot

The angle plot compares both loadings and eigenvalues at once and it can be used to understand the information loss if the data structure of one dataset is superimposed on the other, thereby revealing which principal components (i.e. loading and eigenvalue pairs) are most similar and most different across the two datasets. Let $\lambda_{\max} = \max\{\lambda_1^x, \lambda_1^y\}$ be the largest eigenvalue for the two datasets. Then the empirical correlation matrix for the $\mathbf{X}$ dataset, S_x, has the following orthogonal decomposition in the coordinate system of the $\mathbf{Y}$ dataset

$$S_x = \sum_{j=1}^{d} \lambda_j^x \eta_j^x (\eta_j^x)^\top = \lambda_{\max} \sum_{k=1}^{d} \left(\sum_{j=1}^{d} \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} \eta_j^y (\eta_j^{y\top} \eta_k^x) \right) \left(\sum_{j=1}^{d} \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} \eta_j^y (\eta_j^{y\top} \eta_k^x) \right)^\top,$$

and we have a similar decomposition of S_y in the coordinate system of the $\mathbf{X}$ dataset. We propose to visualize these two decompositions in a $d \times d$ grid display. In the jth row and kth column of this display we plot two arrows based at the lower left corner of the grid cell. The first arrow has length μ_{jk} and angle $\theta_{jk}/2$ counterclockwise from the diagonal, and the second arrow has length ν_{jk} and angle $\theta_{jk}/2$ clockwise from the diagonal. To facilitate the following description we will refer to the arrows drawn counterclockwise as the upper arrows, and the arrows drawn clockwise as the lower arrows. The lengths μ_{jk} and ν_{jk}, and the angle θ_{jk}, are given by

$$\mu_{jk} = \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} |\eta_k^{x\top} \eta_j^y|, \qquad \nu_{jk} = \sqrt{\frac{\lambda_j^y}{\lambda_{\max}}} |\eta_j^{y\top} \eta_k^x|, \qquad \theta_{jk} = \arccos(|\eta_k^{x\top} \eta_j^y|).$$

For two d-dimensional, unit length vectors a and b, it holds that $a^\top b = \langle a, b \rangle = \tilde{c}(a, b)$, where $\tilde{c}$ denotes the empirical correlation function. Thus, in the angle plot, we are essentially looking at the absolute values of correlations between loadings that have been scaled according to their variance contributions. The absolute value of the projection $\eta_k^{x\top} \eta_j^y$ is inserted due to the indeterminacy of the direction of loading vectors. This indeterminacy implies that the angle between loadings from the two datasets can always be chosen to be in the interval $[0, \pi/2]$, and hence the decomposition of S_x and S_y can be visualized in a joint plot by dividing the angles by two and using counterclockwise and clockwise shifts from the diagonal. Furthermore, the scaling of the lengths by $\lambda_{\max}$ is made so that the longest arrow has at most unit length.

In the *angle plot*, the upper arrows in the kth column of the grid display visualize the decomposition of the kth principal component of the A dataset in the coordinate system of the B dataset. Similarly, the lower arrows in the jth row visualize the decomposition of the jth principal components of the B dataset in the coordinate system of the A dataset. If we have coinciding upper and lower arrows along the diagonal in the grid display, and arrows of vanishing lengths in the off-diagonal cells, then the structures of the two datasets are identical. In contrast, differences in the eigenvalues are visualized as differences in the lengths of the upper and the lower arrows, also in the diagonal. And loadings in other directions than the corresponding loading from the other dataset are visualized as angle separation of the upper and the lower arrows in the diagonal cells, as well as arrows of non vanishing length in the off-diagonal cells.

However, there is a generic situation where the angles between the arrows lack discriminative power, namely when two or more neighboring eigenvalues are identical or close. We refer to this situation as *eigenvalue indeterminancy*. For instance, multiple eigenvalues close to 1 arise if more than one of the variables are uncorrelated, or close to uncorrelated, of all the other variables. In this situation only the corresponding eigenspace is well estimated, but the rotation of the associated eigenvectors will be subject to sampling variability. To investigate if we are in such a situation of identical data structures with neighbouring eigenvalues being close, where comparably long arrows may occur in the corresponding blocks of off-diagonal cells, we have implemented a permutation test similar as done for the CE-plot. For each random reallocation we recompute the angles θ_{jk}. The resampled angles may both be used for quantitative tests, where large angles are critical for the diagonal cells and small angles are critical for the off-diagonal cells, as well as for the visualisation of the sampled angle regions. In the visualization we propose to present the entire sampled region of angles for each cell, but gray shaded according to percentiles such that more critical angles are lighter gray.

The chroma plot

The chroma plot is primarily an illustration of differences in loading patterns and it targets the question of how the roles of the original variables are different between the two datasets, thus leading the data structure comparison question back to its original, empirical context. The chroma plot consists of two panels, one for each dataset, made up of colored bars. These bars each represent a principal component and their coloring illustrates the relative weights of the d original variables, that is, their absolute, normalized loading contributions. More specifically, when illustrating the ith principal component in the two data subsets, we plot vertical bars of length one that has been divided into d segments of different colors. The width of the jth colored segment is given by

$$\omega_{ij}^x = \frac{|\eta_{ij}^x|}{\sum_{k=1}^d |\eta_{ik}^x|}, \qquad \omega_{ij}^y = \frac{|\eta_{ij}^y|}{\sum_{k=1}^d |\eta_{ik}^y|},$$

where η_{ij}^x and η_{ij}^y denotes the jth entry of η_i^x and η_i^y, respectively. Due to the indeterminacy of the loading signs, all the signs are removed from the coefficients in the loadings. The bars are ordered according to the eigenvalues and they are annotated with the cumulative percentage explained variance of that component, that is, the scaled and summed variance contributions,

$$\tilde{\Sigma}_i^x = \frac{\sum_{j=1}^i \lambda_j^x}{\sum_{k=1}^d \lambda_k^x} = \frac{\sum_{j=1}^i \lambda_j^x}{d}, \qquad \tilde{\Sigma}_i^y = \frac{\sum_{j=1}^i \lambda_j^y}{\sum_{k=1}^d \lambda_k^y} = \frac{\sum_{j=1}^i \lambda_j^y}{d}.$$

Especially when d is large, we recommend plotting only a select set of interesting principal components (for instance identified by use of the angle plot). In this scenario, the annotations should rather be the non-cumulative variance contributions, $\tilde{\sigma}_i^x = \frac{\lambda_i^x}{d}$ and $\tilde{\sigma}_i^y = \frac{\lambda_i^y}{d}$. If there is eigenvalue indeterminancy for neighboring components, as discussed above, the loading patterns for these components will not have a meaningful interpretation.

Therefore, one should not compare the chroma bars for component that were identified to suffer from eigenvalue indeterminancy by the angle plot.

The plots resulting from this procedure should be inspected focusing on two properties: Similarities in loading patterns, which will correspond to similar visual impressions, and similarities in variance contributions. For each component, the loadings describe how influential the different variables are on that component. Therefore, the chroma plot allows us to make qualitative statements about the original datasets, such as *"variable 1 is generally more influential in subset A than it is in subset B"*, thereby helping us to understand where and why potential data structure differences are found.

Further considerations

Though the PCADSC toolbox comprises only three diagnostic plots as of now, the PCA decomposition can of course be utilized further to bring about new PCADSC plots that illustrate more aspects of data structure differences. For instance, the scree plot (as introduced by Cattell [2]), which plots the ordered eigenvalues against their component numbers, could easily be recreated as a PCADSC tool by creating scree plots for both data subsets in the same coordinate system, yielding a simple graphical comparison. The resampling approach used in the CE plot could then be adopted, thus annotating the scree plots with resampled curves. This general approach to model-fit evaluation is very helpful in combination with permutation tests, as inspired by Lin and colleagues [5].

The proposed methods have a few limiations that should be mentioned. First of all, as PCA is performed after standardization of the variables, the PCADSC methods cannot detect differences in neither mean values nor in marginal variances - this information is thrown away as the very first step. However, such differences are of a fundamentally different nature than those we have discussed so far. Differences in mean values are often the main interest of the analysis and should of course not be regarded as a data comparability issue. Differences in marginal variances, on the other hand, can pose some modeling challenges, but none that are not generally solvable by use of random effects models. We therefore do not consider this property of PCADSC to be a major drawback.

A more prominent limitation is the fact that the PCASDC methods are only valid for variables that have a numerical interpretation. In particular, this means that the methods cannot be used for data with nominal categorical variables. This limitation is inherited from the PCA method and thus, moving beyond it will entail replacing the PCA framework with a more general one. Multiple correspondence analysis (MCA) is often suggested as a categorical extension of PCA [1] and including MCA in the PCADSC tools is thus a natural next step.

Not all of the PCADSC tools scale well with a growing number of variables, d. While the CE plot scales well with large d, the angle plot and the chroma plot can become difficult to read if there are a lot of variables and thus a large number of principal components. However, because the angle plot lets the arrow length depend on the size of the correpsonding eigenvalue, its visual impression should help guide a user towards influential *and* dissimilar components even though d is large. And when these components have been identified, the chroma plot can focus only on them, thereby avoiding plotting the loading patterns of too many components.

References

[1] H. Abdi and L.J. Williams, *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics 2 (2010), pp. 433–459. Available at http://dx.doi.org/10.1002/wics.101.

[2] R.B. Cattell, *The Scree Test For The Number Of Factors*, Multivariate Behavioral Research 1 (1966), pp. 245–276.

[3] J. Helliwell, R. Layard, and J. Sachs, *World Happiness Report 2016 Update*, New York: Sustainable Development Solutions Network (2016).

[4] K. Jeffrey, S. Abdallah, and A. Quick, *Europeans' Personal and Social Wellbeing: Topline Results from Round 6 of the European Social Survey*, ESS Topline Results (Series 5) (2015).

[5] D.Y. Lin, L.J. Wei, and Z. Ying, *Model-Checking Techniques Based on Cumulative Residuals*, Biometrics 58 (2002), pp. 1–12.

[6] H.L. Lolle and J.G. Andersen, *Measuring Happiness and Overall Life Satisfaction: A Danish Survey Experiment on the Impact of Language and Translation Problems*, Journal of Happiness Studies 17 (2016), pp. 1337–1350. Available at http://dx.doi.org/10.1007/s10902-015-9646-4.

[7] A.H. Petersen and B. Markussen, *PCADSC: Tools for Principal Component Analysis-Based Data Structure Comparisons* (2017). Available at https://CRAN.R-project.org/package=PCADSC, R package version 0.8.0.

[8] A. Rodríguez-Pose and V. von Berlepsch, *Social Capital and Individual Happiness in Europe*, Journal of Happiness Studies 15 (2014), pp. 357–386.

[9] R. Veenhoven, *Cross-national differences in happiness: Cultural measurement bias or effect of culture?*, International Journal of Wellbeing 2 (2012), pp. 333–353.

PCA-based exploratory data structure comparisons – or *A Story of Happiness Rankings*

Anne Helby Petersen[1]*, Bo Markussen[2] & Karl Bang Christensen[1]

[1]Department of Public Health, University of Copenhagen; [2]Department of Mathematical Sciences, University of Copenhagen

Abstract: Many datasets are somehow divided into distinct subsets, e.g. due to multi-center sampling, survey experiments or multiple measurement instruments being applied. However, we often still wish to analyze the data jointly, and therefore, we should consider whether the data subsets are sufficiently similar in structure for this endeavor to be meaningful. We propose to address this question of data structure similarity by use of three new diagnostic plots, all based on principal component analysis. We apply these exploratory methods to data from the European Social Survey measuring psychological well-being and discuss whether or not this concept is comparable across countries, as is often done e.g. in the UN Happiness Report, where Denmark is always in the top. Are we really happier than everybody else, or could it be that the UN happiness definition is just designed to measure our kind of happiness?

1 Introduction

Data comparability is a recurring topic in applied statistics. Often, data are collected in such a way that it is essentially divided into several subsets whose comparability needs to be assessed empirically. This happens for instance when data are collected across several centers (or countries) or when different versions of an instrument, the mode of administration of a survey, or the order of survey questionnaire items are applied. An example of where the data structure comparability issue must be considered is when conducting international rankings of countries according to e.g. educational quality (e.g. the PISA project) and citizen happiness (e.g. the UN *World Happiness Report* project). From a methodological point of view, such international rankings are problematic, as they rely on the fundamental assumption that the measured concepts are inherently the same across countries.

The comparability issue in these examples can be summarized as follows: Assume that we have two datasets with the same variables, but different observations, often represented as a single dataset with a subset-inducing variable, and that we wish to compare them without specifying a model, or even a variable of interest. The central question is then whether the two datasets can readily be combined for the purpose of later data analysis, or if the subset-inducing variable implies heterogeneity that must be dealt with in later statistical modelling. Sophisticated methods for addressing this question are available when we are willing to assume a statistical model. However, this places the effort of assessing data comparability very late in the data analysis process and it makes the comparability assessment ad-hoc, as it essentially relies on the appropriateness of the modeling choices, which again depends on the structure of the data. We propose three diagnostic plots that can be used for

*Corresponding author: ahpe@sund.ku.dk

data structure comparisons without having to specify a statistical model. These methods use principal component decomposition of the empirical covariance matrix in the two subsets to create intuitive visualizations of data structure differences and we refer to them collectively as *Principal Component Analysis-based Data Structure Comparisons* (PCADSC). Note that the use of PCA implies that we can only compare datasets with numerical variables and that we focus on the linear aspect of the data structures.

We will showcase how the plots can help detect and understand data structure differences by applying them to data on psychological well-being from three European countries that may or may not be comparable. Our starting point is Denmark which has repeatedly been awarded with the title of "happiest country in the world" by the *World Happiness Report*, most recently in 2016 [3], and we wish to investigate if this title is really meaningful at all. In the rankings of happiness, not much work has yet been devoted to evaluating the assumption of international comparability, though Veenhoven presents a theoretically thorough, but empirically simplistic, summary of possible reasons for lack of comparability [9], while Lolle & Andersen show highly potent translation issues for the term *happiness* [6]. We will address the question by focusing on a single aspect of happiness, namely psychological well-being and we will compare Denmark with Bulgaria and Sweden. Though both Denmark and Bulgaria are European countries and thus neither very far apart geographically nor culturally, these two countries have previously been highlighted to be very different in terms of what defines happiness [4]. Moreover, intra-European regional differences in the relationship between social capital and happiness have also been demonstrated; when comparing Northern European countries to other European countries, a much less pronounced relationship between the two concepts is generally found [8]. In particular, interpersonal relations play a less important role in Denmark, compared to Bulgaria. Therefore, a successful method for data comparisons should be able to detect these differences by looking at data on psychological well-being from these two countries. We also compare the Danish data with Swedish data in order to investigate if the PCADSC tools actually do hold discriminatory power. Denmark and Sweden are often deemed very similar in terms of culture and history. Therefore, we expect fundamental concepts, such as psychological well-being, to be similar across these two countries.

2 Methods

We will use the three diagnostic PCADSC plots to compare the covariance matrices of the Danish and Bulgarian data, and the Danish and Swedish data, respectively. Note that if all variables in the subsets are jointly normal with known means, covariance matrices are sufficient statistics describing the joint distribution of all the variables, but even without the normality assumption, pairwise correlations and marginal variable variances are still interesting quantities describing linear interrelations between variables. This makes the empirical covariance matrix a reasonable place to start looking for differences in data structures. Computing empirical covariance matrices for the two subsets and comparing them entry-by-entry becomes increasingly difficult as the number of variables increases and for this reason we decompose and recompose the covariance matrices using *principal component analysis* (PCA) to get an overview of differences between the two subsets.

We propose the following three diagnostic plots:

The CE plot: The CE (cumulative eigenvalue) plot can be used to illustrate differences in eigenvalues, that is, in the relative importance of the directions identified by the PCA. The CE plot is accompanied by two permutation-tests (using the *Kolmogorov-Smirnov* and the *Cramér-von Mises* test statistics, respectively), testing the hypothesis of no difference in variance components. This plot displays a cummulative eigenvalue curve together with a 95% confidence region constructed under the null-hypothesis of no difference in eigenvalues. A CE curve that

falls within the confidence region thus suggests similar eigenvalues.

The angle plot: The angle plot compares both loadings and eigenvalues at once and it can be used to understand the information loss if the data structure of one dataset is superimposed on the other, thereby revealing which principal components (i.e. loading and variance component pairs) that are most similar and most different across the two datasets. The plot displays an angle for each pair of loadings from the two datasets under comparison, and the size of the angle is proportional to the correlation between the relevant loadings. The lengths of the lines represent the sizes of the eigenvalues, which means that angles with longer lines correspond to components that explain more variance. The angles are accompanied by shaded regions displaying the distribution of the angles under the null-hypothesis of no difference in data structures. Angles that fall within the shaded regions thus support a conclusion of similar data structures.

The chroma plot: The chroma plot is primarily an illustration of the loading patterns and it targets the question of how the roles of the original variables are different between the two datasets, thus leading the data structure comparison question back to its original, empirical context. Each component of interest is illustrated with a horizontal bar, divided into blocks. Each block represents a variable and the size of the block represents its weight in the loading vector of the principal component. Thus, similar overall visual impressions will imply similar loading structures.

Further details and a full mathematical specification of the PCADSC plots are available in Appendix A. We suggest that the plots are applied in the order in which they were presented, thereby moving from overall assesments of similarity to explanations about the nature of potential data structure differences. This workflow is summarized in Figure 1. The plots are all implemented in the R package `PCADSC` [7], which is available on CRAN, and can be tried out very easily using only a few lines of code:

```
library(PCADSC)
results <- PCADSC(data, splitBy = "category")
CEPlot(results) #make CE plot
anglePlot(results) #make angle plot
chromaPlot(results) #make chroma plot
```

Here, `data` is a dataset where the two subsets to be compared are indicated by a variable named `category`.

3 Comparing psychological well-being across countries

We use data from the 2012 version of the European Social Survey (ESS) project to investigate inter-country differences in psychological well-being and happiness.

3.1 Data

The ESS 2012 data contain a total of 626 variables collected from 54673 citizens of 29 countries. Here, we will only work with a subset of 35 questionnaire items that are all related to psychological well-being. These 35 items can be divided into 6 distinct scales, namely *Evaluative well-being*, *Emotional well-being*, *Functioning*, *Vitality*, *Community well-being*, and *Supportive relationships*. More details on these scales can be found in [4]. We represent each of the scales by a single variable, calculated as the average score on the items related to

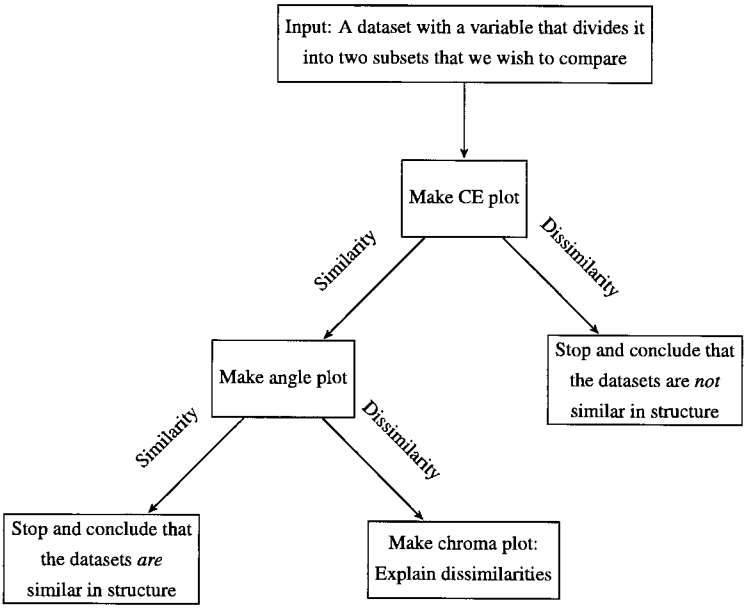


Figure 1: The PCADSC workflow, starting with two datasets and then successively applying the three visual tools for obtaining a deeper and deeper understanding of similarities and dissimilarities in the data structures.

	Denmark	Bulgaria	Sweden
Evaluative well-being	8.75 (8.00, 9.50)	5.00 (3.50, 7.00)	8.00 (7.00, 9.00)
Emotional well-being	8.33 (7.22, 8.89)	6.67 (5.00, 7.78)	7.78 (6.67, 8.89)
Functioning	7.57 (6.93, 8.21)	6.68 (5.50, 7.68)	7.04 (6.39, 7.68)
Vitality	7.50 (6.67, 8.33)	7.50 (5.83, 8.33)	7.50 (6.67, 9.17)
Community well-being	6.77 (5.83 ,7.57)	4.67 (3.70, 5.70)	6.57 (5.66, 7.37)
Supportive relationships	8.25 (7.42, 8.92)	7.25 (6.17, 8.08)	8.25 (7.42, 8.75)

Table 1: Median values of each of the six dimensions of psychological well-being, stratified by country. 25 and 75 percentiles are listed in parentheses. The scales are constructed such that they all run from 0-10.

that variable and scaled to take values between 0 and 10. For simplicity, we use only complete cases for this construction and thus exclude all participants that did not answer all the 35 questionnaire items used below. This corresponds to approximately 9% of the observations in the Danish sample, 20% in the Bulgarian sample and only 6% in the Swedish sample. All in all, we have $n_{DK} = 1498$ complete case observations in the Danish sample, $n_{BG} = 1798$ observations in the Bulgarian sample and $n_{SE} = 1736$ Swedish observations. Table 1 summarizes the marginal distributions of the six dimensions of psychological well-being, stratified by country.

3.2 Comparing Denmark and Bulgaria

Figure 2 presents the CE plot and the angle plot obtained from comparing the Danish and Bulgarian psychological well-being scales. The CE plot show a remarkable degree of lacking comparability: The cumulative differences in the eigenvalues by far exceed what could come about randomly if there really were no difference in the data structures. This is also confirmed by the Kolmogorov-Smirnov and the Cramér-von Mises tests, which both result

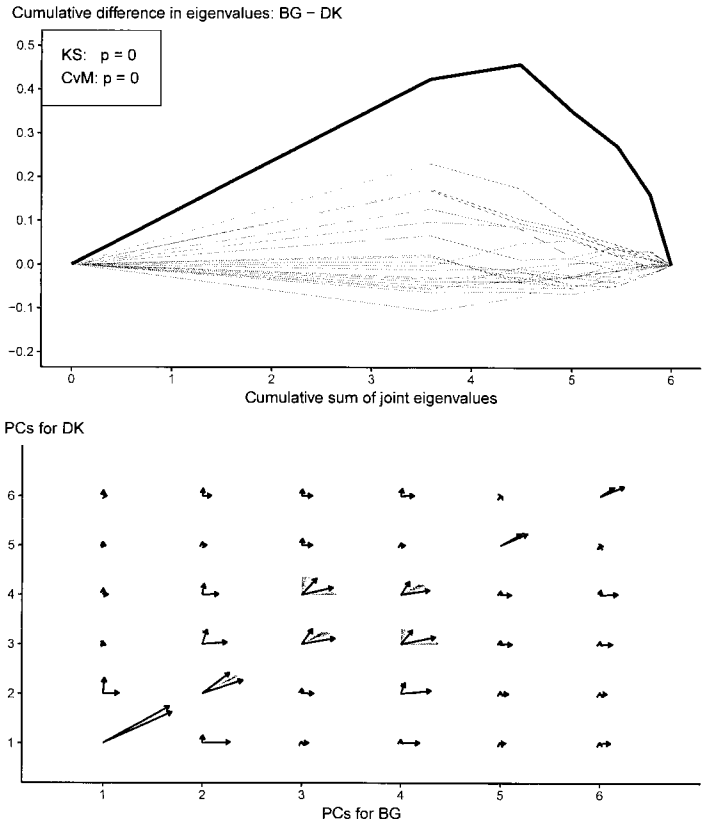


Figure 2: CE plots (top panel) and angle plots (bottom panel) comparing Danish and Bulgarian data on psychological well-being. The CE plot is annotated with the p-values of the Kolmogorov-Smirnov and the Cramér-von Mises tests of the hypothesis of no difference in data structures. In the angle plot, the upper arrows show the principal components of the Bulgarian dataset decomposed in the coordinate system of the principal components of the Danish dataset, while the lower arrows illustrate the reverse.

in p-values that are virtually zero. The CE curve generally lies above zero, suggesting that the Bulgarian dataset has larger eigenvalues for all but the last component, thereby explaining more variance at lower dimensionalities.

Moving on to the angle plot, we find that the differences are primarily to be found in the second, third, and fourth principal components (PCs). The upper arrows visualize the decomposition of the principal components for the Bulgarian dataset in the coordinate system of the Danish dataset. We see that PC2 also loads on PC3, that PC3 also loads on PC4, and that PC4 also loads on PC2 and PC3. The lower arrows visualize the decomposition of the principal components for the Danish dataset in the coordinate system of the Bulgarian dataset. Here, we see that PC2 also loads on PC4, that PC3 also loads on PC2 and PC4, and that PC4 also loads on PC3. Thus, if we wish to understand why differences in the data structures occur, an inspection of the loadings of components 2, 3 and 4 might be informative.

The chroma plot in Figure 3 allows us to look closer into these components. Here, we find that the relative importance of the *Community well-being* and *Supportive relationships* scales is much larger in the Bulgarian sample than in the Danish. In the Danish data, on the other hand, we find that *Vitality* and *Emotional well-being*

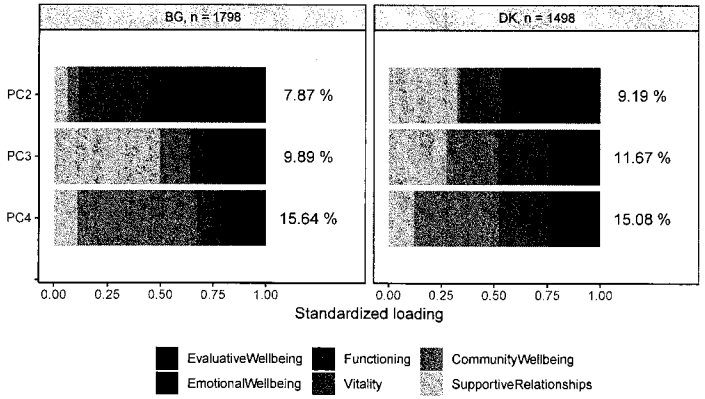


Figure 3: Chroma plot for comparing Danish and Bulgarian data on psychological well-being. A chroma plot comparing the 2nd, 3rd, and 4th principal components of the Bulgarian- and Danish psychological well-being data. The component-bars are annotated with their relative variance contributions.

seem to play bigger roles, as they appear with larger loadings in more high-ranking components in this sample, relative to the Bulgarian.

All in all, we find that psychological well-being does *not* seem to be the same concept in Bulgaria and Denmark. The two countries disagree both in how many dimensions are needed to capture the most important parts of the concept (as illustrated by the differences in eigenvalues) and in how these dimensions are then weighted among the 6 scales (as illustrated by the angle- and chroma plots). In Bulgaria, interpersonal features seem to be more informative of psychological well-being, whereas in Denmark, individual characteristics play a relatively larger role, corresponding well with the previous findings mentioned above. Thus, the datasets are fundamentally different and we should therefore be wary about combining them in a joint analysis, which was also the conclusion of the ESS authors, though based on country-level aggregated statistics [4]. Moreover, this also implies that two countries cannot be ranked in terms of which country is "the most happy", at least not by referring to psychological well-being dimensions such as those encountered here.

3.3 Comparing Denmark and Sweden

We now turn to the comparison of Denmark and Sweden in terms of psychological well-being. Figure 4 shows the CE- and angle plots for these two countries. In the CE plot, we now find the cumulative eigenvalue curve to be just within the acceptance region of the null-hypothesis. This is also reflected by the two tests, which now produce p-values of $p_{\text{KS}} = 0.17$ and $p_{\text{CvM}} = 0.11$, respectively. Thus, it is not unreasonable to assume equal eigenvalues in the two datasets.

The angle plot in Figure 4 shows that the two datasets agree very strongly about the relative importance of the six scales in the six PCs, as almost all off-diagonal arrows are practically non-existent. This implies that if one already has the information held in the first PC from the Danish data, this information is in itself mostly sufficient to describe the first PC of the Swedish data.

Looking at the chroma plot in Figure 5, the same tale is told once again: Here, we find remarkably similar loading patterns in the first three components (which are responsible for almost 80% of the variance in both datasets), and slight, but increasing, differences in the remaining three components. We therefore conclude that

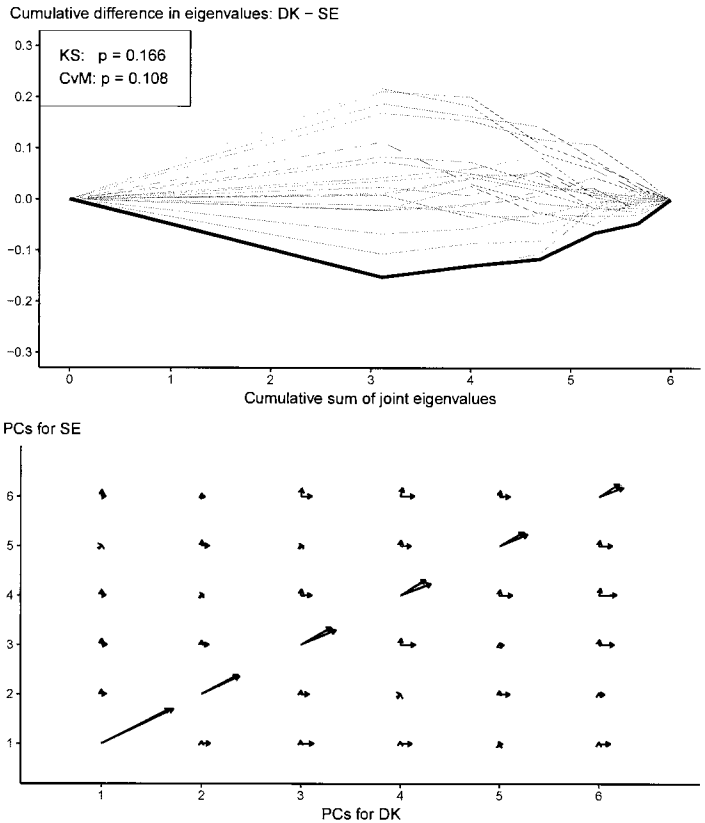


Figure 4: A CE plot (top panel) and an angle plot (bottom panel) comparing Danish and Swedish data on psychological well-being. The upper arrows show the principal components of the Danish dataset decomposed in the coordinate system of the principal components of the Swedish dataset, and the lower arrows illustrate the reverse.

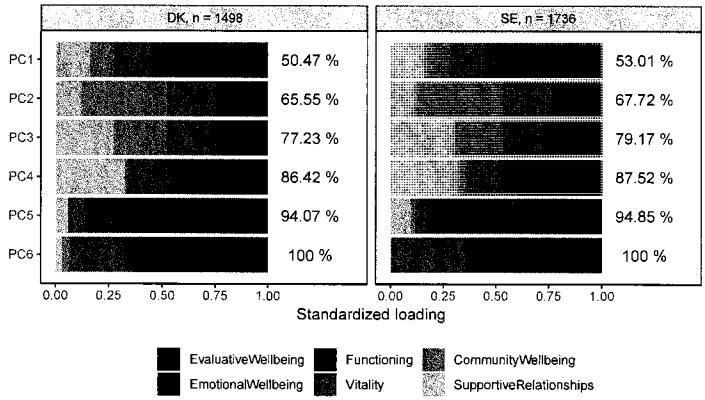


Figure 5: Chroma plot comparing the loading patterns of the Danish and the Swedish subsamples. For each component, the bar is annotated with its cumulative variance contribution, that is, how much variance can be explained by having information of this and the preceding components.

any differences in the data structures of the Danish and the Swedish samples are related to the least important dimensions of the datasets and that these dimensions are only responsible for less than 25% of the variance in both datasets. In particular, this means that we can combine and compare the Danish and Swedish datasets in a meaningful way and for instance conclude using Table 1 that in general, Danes seem to be somewhat more happy than Swedes, and in particular that the least happy people in Denmark (represented by the 1st quartiles) are generally happier than the least happy people in Sweden. A more thorough, statistical investigation could now be put to work on answering *why* this seems to be the case.

4 Conclusion

We have presented three new tools for visual, exploratory data structure comparisons that utilizes covariance matrix decompositions obtained by use of principal component analysis, and used them to investigate whether comparisons of psychological well-being scores are meaningful across Denmark, Bulgaria and Sweden. We found that while Denmark and Sweden seem to have similar concepts of psychological well-being, Denmark and Bulgaria do not. This means that international rankings of psychological well-being are not justified. And this may further imply that rankings according to more broad concepts that rely on psychological well-being, such as *happiness*, may not be meaningful.

Appendix A: Mathematical details of PCADSC

The tools presented here are all based on comparing the PCA results across two different datasets that contain the same variables. In our approach, we standardize the variables before conduction the PCA, which is equivalent to working with the empirical correlation matrices rather than the empirical covariance matrices. We also assume that all considered variables have a numerical interpretation, for example by being continuous or ordinally categorical. Moreover, we use the following notation: $\mathbf{X} \in \mathbb{R}^{n_x \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n_y \times d}$ are datasets containing the same number of variables, d, but possibly different numbers of observations, n_x and n_y. We use X_j to refer to the

jth variables of $\mathbf{X}$ and x_{ij} to refer to the ith observation within that variable, while $x_{i\cdot}$ is the full ith observation row. Note that $X_j \in \mathbb{R}^{n_x}$, $x_{ij} \in \mathbb{R}$, and $x_{i\cdot} \in \mathbb{R}^d$. We let $\bar{X} = (\bar{X}_1, ..., \bar{X}_d)^T = \frac{1}{n_x}\sum_{i=1} x_{i\cdot}$ denote the variable averages and use $S_x = \frac{1}{n_x - 1}\sum_{i=i}^{n}(x_{i\cdot} - \bar{X})(x_{i\cdot} - \bar{X})^T \in \mathbb{R}^{d\times d}$ to denote the empirical covariance matrix of $\mathbf{X}$. We let

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \in \mathbb{R}^{(n_x+n_y)\times d}$$

denote the combined dataset.

For each of the three datasets, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$, we complete the following steps (here described for $\mathbf{X}$ only):

1. Standardize each of the variables to have mean zero and unit standard deviation. Let $\tilde{\mathbf{X}}_j \in \mathbb{R}^{n_x\times d}$ be the standardized dataset.

2. Form the principal component analyses

$$S_x = \frac{1}{n_x - 1}\sum_{i=1}^{n_x} \tilde{x}_{i\cdot}\tilde{x}_{i\cdot}^\top = \sum_{j=1}^{d} \lambda_j^x \eta_j^x (\eta_j^x)^\top$$

   thereby obtaining loadings η_j^x and eigenvalues λ_j^x for $j = 1, \ldots, d$.

The hereby obtained PCA decompositions of the correlation matrices can then be compared. The standardization implies that the diagonal elements of S_x, S_y, and S_z all equal 1, and thus also that

$$\sum_{j=1}^{d} \lambda_j^x = \sum_{j=1}^{d} \lambda_j^y = \sum_{j=1}^{d} \lambda_j^z = d.$$

This identity will simplify some expressions below.

The cumulative eigenvalue plot

The cumulative eigenvalue (CE) plot compares the eigenvalues of the correlation matrices. These eigenvalues represent how much information is withheld in each component in terms of explained variance. Thus, by comparing cumulative sums of eigenvalues, it is possible to obtain a detailed picture of how the two datasets differ in terms of what components are the most informative. In order to investigate whether the same proportion of the total variation can be described by the same number of principal components in the two datasets, we plot a piecewise linear curve connecting the points

$$(0,0),\ (\lambda_1^z, \lambda_1^x - \lambda_1^y),\ (\lambda_1^z + \lambda_2^z, \lambda_1^x + \lambda_2^x - \lambda_1^y - \lambda_2^y),\ \ldots,$$
$$\left(\sum_{j=1}^{d} \lambda_j^z, \sum_{j=1}^{d} \lambda_j^x - \sum_{j=1}^{d} \lambda_j^y\right)$$

Due to the standardization, the last point will always be equal to $(d, 0)$. Thus, this curve will begin and end at the x-axis. And the larger excursions it makes away from the first axis, the less alike the cumulative eigenvalue sums for the two datasets are. Moreover, a positive cumulative difference implies that dataset $\mathbf{X}$ holds more information in the first components than dataset $\mathbf{Y}$ does. By using the cumulative eigenvalues of $\mathbf{Z}$ as the first coordinate, we obtain a visualization where a component takes up an amount of horizontal space that corresponds to the variance explained by the component (that is, its eigenvalue), thereby letting the more influential components drive the visual impression.

In order to test whether these cumulative differences are statistical artefacts or if they represent something real, we have implemented both the *Kolmogorov-Smirnov* and the *Cramér-von Mises* test statistics, which are given by

$$\mathrm{KS} = \max_{k=1,\ldots,d} \left| \sum_{j=1}^{k} \lambda_j^x - \sum_{j=1}^{k} \lambda_j^y \right|, \qquad \mathrm{CvM} = \sum_{k=1}^{d-1} \frac{\lambda_k^z + \lambda_{k+1}^z}{2} \left(\sum_{j=1}^{k} \lambda_j^x - \sum_{j=1}^{k} \lambda_j^y \right)^2.$$

We conduct the tests as *permutation tests*, that is, by randomly reallocating the combined and separately standardized datasets into two new datasets of n_x and n_y observations, respectively, and then redoing the CE plot steps and recalculating the test statistics. This should be done a large (e.g. 10000) number of times. Then, a p-value is obtained by computing the proportion of reallocated datasets that lead to test statistics at least as large as the one we found for the original datasets.

The permutation test results are also used to visualize the uncertainty of the CE curve in the plots. In the CE plots shown in the following section, we plot the observed curve together with 20 of the resampled curves, as well as a shaded region visualizing pointwise 95 % coverage intervals. If the observed curve is very different from the resampled curves or if it is substantially outside the shaded region, then this also indicates differences between the two datasets.

The angle plot

The angle plot compares both loadings and eigenvalues at once and it can be used to understand the information loss if the data structure of one dataset is superimposed on the other, thereby revealing which principal components (i.e. loading and eigenvalue pairs) are most similar and most different across the two datasets. Let $\lambda_{\max} = \max\{\lambda_1^x, \lambda_1^y\}$ be the largest eigenvalue for the two datasets. Then the empirical correlation matrix for the $\mathbf{X}$ dataset, S_x, has the following orthogonal decomposition in the coordinate system of the $\mathbf{Y}$ dataset

$$S_x = \sum_{j=1}^{d} \lambda_j^x \eta_j^x (\eta_j^x)^\top = \lambda_{\max} \sum_{k=1}^{d} \left(\sum_{j=1}^{d} \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} \eta_j^y (\eta_j^{y\top} \eta_k^x) \right) \left(\sum_{j=1}^{d} \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} \eta_j^y (\eta_j^{y\top} \eta_k^x) \right)^\top,$$

and we have a similar decomposition of S_y in the coordinate system of the $\mathbf{X}$ dataset. We propose to visualize these two decompositions in a $d \times d$ grid display. In the jth row and kth column of this display we plot two arrows based at the lower left corner of the grid cell. The first arrow has length μ_{jk} and angle $\theta_{jk}/2$ counterclockwise from the diagonal, and the second arrow has length ν_{jk} and angle $\theta_{jk}/2$ clockwise from the diagonal. To facilitate the following description we will refer to the arrows drawn counterclockwise as the upper arrows, and the arrows drawn clockwise as the lower arrows. The lengths μ_{jk} and ν_{jk}, and the angle θ_{jk}, are given by

$$\mu_{jk} = \sqrt{\frac{\lambda_k^x}{\lambda_{\max}}} |\eta_k^{x\top} \eta_j^y|, \qquad \nu_{jk} = \sqrt{\frac{\lambda_j^y}{\lambda_{\max}}} |\eta_j^{y\top} \eta_k^x|, \qquad \theta_{jk} = \arccos(|\eta_k^{x\top} \eta_j^y|).$$

For two d-dimensional, unit length vectors a and b, it holds that $a^\top b = \langle a, b \rangle = \tilde{c}(a, b)$, where $\tilde{c}$ denotes the empirical correlation function. Thus, in the angle plot, we are essentially looking at the absolute values of correlations between loadings that have been scaled according to their variance contributions. The absolute value of the projection $\eta_k^{x\top} \eta_j^y$ is inserted due to the indeterminacy of the direction of loading vectors. This indeterminacy implies that the angle between loadings from the two datasets can always be chosen to be in the interval $[0, \pi/2]$, and hence the decomposition of S_x and S_y can be visualized in a joint plot by dividing the angles by two and using counterclockwise and clockwise shifts from the diagonal. Furthermore, the scaling of the lengths by $\lambda_{\max}$ is made so that the longest arrow has at most unit length.

In the *angle plot*, the upper arrows in the kth column of the grid display visualize the decomposition of the kth principal component of the **A** dataset in the coordinate system of the **B** dataset. Similarly, the lower arrows in the jth row visualize the decomposition of the jth principal components of the **B** dataset in the coordinate system of the **A** dataset. If we have coinciding upper and lower arrows along the diagonal in the grid display, and arrows of vanishing lengths in the off-diagonal cells, then the structures of the two datasets are identical. In contrast, differences in the eigenvalues are visualized as differences in the lengths of the upper and the lower arrows, also in the diagonal. And loadings in other directions than the corresponding loading from the other dataset are visualized as angle separation of the upper and the lower arrows in the diagonal cells, as well as arrows of non vanishing length in the off-diagonal cells.

However, there is a generic situation where the angles between the arrows lack discriminative power, namely when two or more neighboring eigenvalues are identical or close. We refer to this situation as *eigenvalue indeterminancy*. For instance, multiple eigenvalues close to 1 arise if more than one of the variables are uncorrelated, or close to uncorrelated, of all the other variables. In this situation only the corresponding eigenspace is well estimated, but the rotation of the associated eigenvectors will be subject to sampling variability. To investigate if we are in such a situation of identical data structures with neighbouring eigenvalues being close, where comparably long arrows may occur in the corresponding blocks of off-diagonal cells, we have implemented a permutation test similar as done for the CE-plot. For each random reallocation we recompute the angles θ_{jk}. The resampled angles may both be used for quantitative tests, where large angles are critical for the diagonal cells and small angles are critical for the off-diagonal cells, as well as for the visualisation of the sampled angle regions. In the visualization we propose to present the entire sampled region of angles for each cell, but gray shaded according to percentiles such that more critical angles are lighter gray.

The chroma plot

The chroma plot is primarily an illustration of differences in loading patterns and it targets the question of how the roles of the original variables are different between the two datasets, thus leading the data structure comparison question back to its original, empirical context. The chroma plot consists of two panels, one for each dataset, made up of colored bars. These bars each represent a principal component and their coloring illustrates the relative weights of the d original variables, that is, their absolute, normalized loading contributions. More specifically, when illustrating the ith principal component in the two data subsets, we plot vertical bars of length one that has been divided into d segments of different colors. The width of the jth colored segment is given by

$$\omega_{ij}^x = \frac{|\eta_{ij}^x|}{\sum_{k=1}^d |\eta_{ik}^x|}, \qquad \omega_{ij}^y = \frac{|\eta_{ij}^y|}{\sum_{k=1}^d |\eta_{ik}^y|},$$

where η_{ij}^x and η_{ij}^y denotes the jth entry of η_i^x and η_i^y, respectively. Due to the indeterminacy of the loading signs, all the signs are removed from the coefficients in the loadings. The bars are ordered according to the eigenvalues and they are annotated with the cumulative percentage explained variance of that component, that is, the scaled and summed variance contributions,

$$\tilde{\Sigma}_i^x = \frac{\sum_{j=1}^i \lambda_j^x}{\sum_{k=1}^d \lambda_k^x} = \frac{\sum_{j=1}^i \lambda_j^x}{d}, \qquad \tilde{\Sigma}_i^y = \frac{\sum_{j=1}^i \lambda_j^y}{\sum_{k=1}^d \lambda_k^y} = \frac{\sum_{j=1}^i \lambda_j^y}{d}.$$

Especially when d is large, we recommend plotting only a select set of interesting principal components (for instance identified by use of the angle plot). In this scenario, the annotations should rather be the non-cumulative variance contributions, $\tilde{\sigma}_i^x = \frac{\lambda_i^x}{d}$ and $\tilde{\sigma}_i^y = \frac{\lambda_i^y}{d}$. If there is eigenvalue indeterminancy for neighboring components, as discussed above, the loading patterns for these components will not have a meaningful interpretation.

Therefore, one should not compare the chroma bars for component that were identified to suffer from eigenvalue indeterminancy by the angle plot.

The plots resulting from this procedure should be inspected focusing on two properties: Similarities in loading patterns, which will correspond to similar visual impressions, and similarities in variance contributions. For each component, the loadings describe how influential the different variables are on that component. Therefore, the chroma plot allows us to make qualitative statements about the original datasets, such as *"variable 1 is generally more influential in subset A than it is in subset B"*, thereby helping us to understand where and why potential data structure differences are found.

Further considerations

Though the PCADSC toolbox comprises only three diagnostic plots as of now, the PCA decomposition can of course be utilized further to bring about new PCADSC plots that illustrate more aspects of data structure differences. For instance, the scree plot (as introduced by Cattell [2]), which plots the ordered eigenvalues against their component numbers, could easily be recreated as a PCADSC tool by creating scree plots for both data subsets in the same coordinate system, yielding a simple graphical comparison. The resampling approach used in the CE plot could then be adopted, thus annotating the scree plots with resampled curves. This general approach to model-fit evaluation is very helpful in combination with permutation tests, as inspired by Lin and colleagues [5].

The proposed methods have a few limiations that should be mentioned. First of all, as PCA is performed after standardization of the variables, the PCADSC methods cannot detect differences in neither mean values nor in marginal variances - this information is thrown away as the very first step. However, such differences are of a fundamentally different nature than those we have discussed so far. Differences in mean values are often the main interest of the analysis and should of course not be regarded as a data comparability issue. Differences in marginal variances, on the other hand, can pose some modeling challenges, but none that are not generally solvable by use of random effects models. We therefore do not consider this property of PCADSC to be a major drawback.

A more prominent limitation is the fact that the PCASDC methods are only valid for variables that have a numerical interpretation. In particular, this means that the methods cannot be used for data with nominal categorical variables. This limitation is inherited from the PCA method and thus, moving beyond it will entail replacing the PCA framework with a more general one. Multiple correspondence analysis (MCA) is often suggested as a categorical extension of PCA [1] and including MCA in the PCADSC tools is thus a natural next step.

Not all of the PCADSC tools scale well with a growing number of variables, d. While the CE plot scales well with large d, the angle plot and the chroma plot can become difficult to read if there are a lot of variables and thus a large number of principal components. However, because the angle plot lets the arrow length depend on the size of the correpsonding eigenvalue, its visual impression should help guide a user towards influential *and* dissimilar components even though d is large. And when these components have been identified, the chroma plot can focus only on them, thereby avoiding plotting the loading patterns of too many components.

References

[1] H. Abdi and L.J. Williams, *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics 2 (2010), pp. 433–459. Available at http://dx.doi.org/10.1002/wics.101.

[2] R.B. Cattell, *The Scree Test For The Number Of Factors*, Multivariate Behavioral Research 1 (1966), pp. 245–276.

[3] J. Helliwell, R. Layard, and J. Sachs, *World Happiness Report 2016 Update*, New York: Sustainable Development Solutions Network (2016).

[4] K. Jeffrey, S. Abdallah, and A. Quick, *Europeans' Personal and Social Wellbeing: Topline Results from Round 6 of the European Social Survey*, ESS Topline Results (Series 5) (2015).

[5] D.Y. Lin, L.J. Wei, and Z. Ying, *Model-Checking Techniques Based on Cumulative Residuals*, Biometrics 58 (2002), pp. 1–12.

[6] H.L. Lolle and J.G. Andersen, *Measuring Happiness and Overall Life Satisfaction: A Danish Survey Experiment on the Impact of Language and Translation Problems*, Journal of Happiness Studies 17 (2016), pp. 1337–1350. Available at http://dx.doi.org/10.1007/s10902-015-9646-4.

[7] A.H. Petersen and B. Markussen, *PCADSC: Tools for Principal Component Analysis-Based Data Structure Comparisons* (2017). Available at https://CRAN.R-project.org/package=PCADSC, R package version 0.8.0.

[8] A. Rodríguez-Pose and V. von Berlepsch, *Social Capital and Individual Happiness in Europe*, Journal of Happiness Studies 15 (2014), pp. 357–386.

[9] R. Veenhoven, *Cross-national differences in happiness: Cultural measurement bias or effect of culture?*, International Journal of Wellbeing 2 (2012), pp. 333–353.

Tildeling af hjemmehjælp til ældre i kommunerne

Niels Henning Bjørn, KL

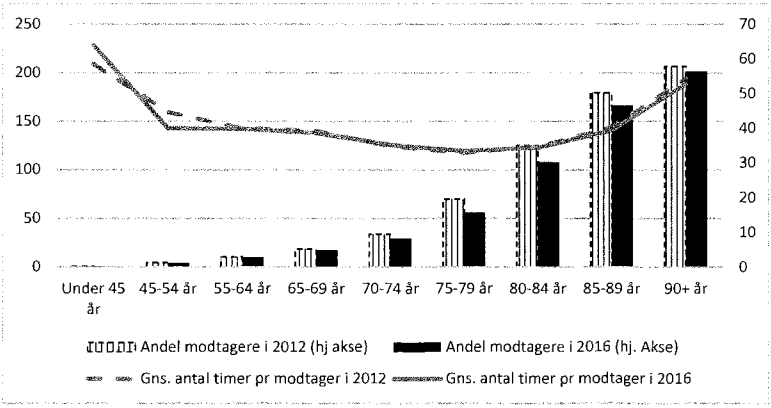
Kommunerne tildeler borgerne i eget hjem hjemmehjælp til praktiske formål og personlig pleje efter behov. Det betyder, at tildelingen af hjemmehjælp er stigende med borgernes alder, mens det modsat gælder, at omfanget af hjemmehjælp er stigende, i takt med at helbredet forringes.

Sammenhængen mellem alder og omfanget af hjemmehjælp bliver meget interessant de kommende år, hvor antallet af ældre over 80 år stiger i takt med at de store årgange født i årene 1942-48 runder de 80 år. Antallet af borgere over 80 år forventes at stige med 60 pct. frem til 2030 og antallet i sidste halvdel af 70'erne forventes at stige med 70 pct. Denne potentielle vækst i behovet for hjemmehjælp er betydelig, og det er nødvendigt at have et godt beslutningsgrundlag for at kunne dimensionere rekrutteringen af personale til hjemmehjælpen, hvorfor det er nødvendigt at kende kriterierne for tilkendelse.

1. Tildeling af hjemmehjælp

Andelen af borgere, som modtager hjemmehjælp i eget hjem, er kraftigt stigende med alderen, og mens 35 pct. af de 80-84 årige modtager hjemmehjælp i 2012, så er det mere end halvdelen af borgerne over 85 år, som modtager hjemmehjælp, jf. figur 1. Antallet af timer pr. modtager af hjemmehjælp vokser derimod ikke så meget ved stigende alder som andelen af modtagere. Dette resultat genfindes også i 2016.

Figur 1: Andel modtagere og gennemsnitligt antal timer af hjemmehjælp

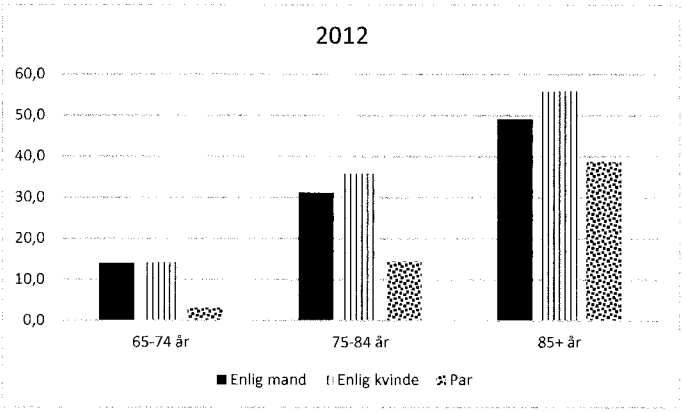


Kilde: Egne beregninger på baggrund af registerdata.

Eksempelvis modtager kun 20 pct. af de 75-79 årige hjemmehjælp, mens det er 50 pct. af de 85-89 årige. Det er en stigning på 150 pct. Det gennemsnitlige antal timer var derimod 118 timer for de 75-79 årige og 144 timer for de 85-89 årig. Forskellen er her på 22 pct. De unge modtagere af hjemmehjælp under 45 år modtager forholdsvis mange timer, men til gengæld modtager kun ganske få i den aldersgruppe hjemmehjælp.

Enlige modtager oftere hjemmehjælp end gifte og samlevende, og det gælder uanset alder, selvom forskellen ikke er så stor for personer over 85 år, jf. figur 2. Kun 14 pct. af borgerne i alderen 75-84 år i parforhold har hjemmehjælp, mens det gælder for mere end 30 pct. af de enlige i samme aldersgruppe.

Figur 2: Andel af borgere, som modtager hjemmehjælp, 2012



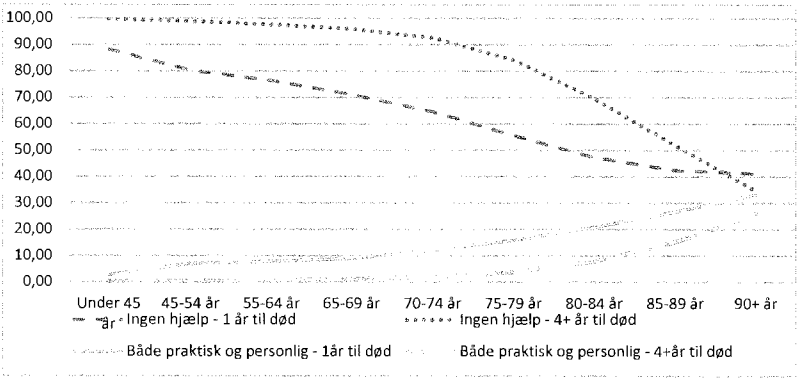
Kilde: Egne beregninger på baggrund af registerdata.

Borgernes helbred spiller en betydelig rolle for tildelingen af hjemmehjælp, men der foreligger kun sjældent oplysninger om helbred og funktionsevne for befolkningen, som kan bruges som grundlag for behovet for hjemmehjælp. I stedet kan bruges en indikator for den forventede tid til død, hvor det antages, at helbred og funktionsevne i gennemsnit forringes frem til dødsdagen. Selvom en person dør inden for en kort tidshorisont, er det dog ikke nødvendigvis ensbetydende med, at personens funktionsevne er så dårlig, at det er nødvendigt at tildele personen hjemmehjælp.

Eksempelvis er det muligt for personer i Danmark d. 1. januar 2012 at finde antallet af døde i en aldersgruppe frem til 1. januar 2013 og de efterfølgende år frem til 1. januar 2018.

Forskellen i andelen uden hjemmehjælp er størst for de 75-79 årige, hvor 84 pct. at dem med mere end fire år til død ikke modtager hjemmehjælp, men det kun gælder for 56 pct. af dem med et år til død, jf. figur 3. Samtidig er bemærkelsesværdigt, at 56 pct. klarer sig uden hjemmehjælp. En forklaring kan dog være, at en andel af dem lever i et parforhold, hvor partneren hjælper i stort omfang.

Figur 3: Andel modtagere og ikke-modtager af hjemmehjælp ift. Til-til-død, 2012



Kilde: Egne beregninger på baggrund af registerdata.

Antallet af modtagere af hjemmehjælp set i forhold til antallet af borgere har været aftagende de seneste, og andelen af personer med hjemmehjælp over 80 år er faldet fra 48 pct. i 2009 til 39 pct. i 2016. Faldet i andelen af modtagere er større for de 65-79 årige, hvor andelen er reduceret fra 13 pct. til 8 pct.

Årsagerne til faldet er formodentlig en kombination af flere faktorer, hvor sund aldring kan forklare en mindre del af, mens de aftalte lofter for det kommunale forbrug i økonomiaftalen mellem Finansministeriet og kommuner lægger grænser for det samlede kommunale forbrug på service. Endelig har kommunerne forøget indsatsen for at forbedre borgernes egne muligheder for at klare sig selv ved at tilbyde genoptræning og rehabilitering.

2. Hvem får hjemmehjælp

Tildelingen af hjemmehjælp kan derfor beskrives ved en række indirekte indikatorer, som skal forklare personernes helbredstilstand, funktionsevne, mulighed for at få hjælp og evne til at klare sig selv. Endvidere inddrages en variabel for hver kommune i datasættet for at tage højde for kommunernes forskellige finansielle situation og prioriteringer.

Tildelingen analyseres ved at estimere i logit for, om borgerne modtager hjemmehjælp, og sammenhængen mellem borgerne og omfanget af hjemmehjælp estimeret ved at bruge en OLS. Derved undgår OLS estimationen problemerne med de mange borgere uden hjemmehjælp.

Udvalget af variabler til estimationen er præget af det-for-hånden-været-søm-princip, idet der ikke foreligger digitaliserede beskrivelse af alle ældre borgeres behov for hjemmehjælp.

Personernes helbredstilstand er heller ikke direkte beskrevet, men forringes generelt med alderen, hvorfor alder kan tages som en indikator på forringelsen af helbredet. Forringelsen kan måske være accelererende, hvorfor alder kvadreret også medtages, jf. Dalgaard og Strulik (2014).

Kønnet inddrages også i estimationerne, da kvinder i gennemsnit lever længere end mænd og må formodes at have et bedre helbred. Omvendt viser de simple opdelinger, at enlige kvinder i alderen 75-84 år oftere modtager hjemmehjælp end enlige mænd. Det understøtter, at kønnet inddrages i beregningerne, hvor det er muligt at inddrage flere faktorer samtidigt.

En anden måde at vurdere helbredet på er at inddrage, hvor mange år den enkelte person lever efter beregningstidspunktet. Bag denne variabel ligger den vurdering, at helbredet og funktionsevnen for personer, der dør inden for en kort årrække, er ringere end helbredet og funktionsevnen for personer, som lever i mange år. Her er det valgt at gennemføre beregningerne i 2012, hvilket gør det muligt at inddele restlevetiden i eksempelvis 0 år, for dem som dør i løbet af året, 1 år for dem som er i live 1. januar 2013, og 5+ år for dem som lever 1. januar 2018 samt alle de mellemliggende år. Ulempen ved denne variabel er, at det ikke er muligt at anvende de aktuelle tal for tildelingen af hjemmehjælp i beregningerne.

Boligens størrelse og ejerforhold kan også være en indikator for personens økonomisk formåen, hvor en skillelinje går mellem ejer- og lejebolig. Her vil en ejerbolig afspejle en større økonomisk formåen og en større bolig vil afspejle en større velstand end en mindre bolig. Tærsklen er her sat ved 85 m^2, så store ejerboliger er 85 m^2 eller større, mens mindre boliger er mindre end 85 m^2.

Civilstanden forventes også at spille en afgørende rolle, hvor enlige oftere får hjemmehjælp end gifte eller andre personer i parforhold i samme aldersgruppe. Dette

kan skyldes, at to personer ved fælles hjælp lettere kan klare de daglige udfordringer som at gøre rent, personlig pleje og de øvrige praktiske udfordringer i hverdagen.

Alternativet kunne være at eventuelle børn kommer og hjælper forældrene eller en enkelt forælder med at klare nogle af de daglige udfordringer. Her er der ikke oplysninger om børnenes hjælp, men det er muligt at finde afstanden mellem forældrenes sogn og sognet, hvor den nærmeste af børnene bor. I begge sogne anvendes midtpunktet af sognet som målepunkt for beregningerne. Hvis de ældre personer ikke har børn eller børnene er flyttet ud af landet, er der ingen til at hjælpe i nærheden, hvilket der også er korrigeret for.

Der er i analysen også taget højde for forskelle mellem kommunerne, hvilket kan skyldes, at kommunerne prioriterer forskelligt, så nogle kommuner prioriterer hjemmehjælp mere eller mindre, så personer med samme funktionsevne og helbred i nogle kommuner får hjemmehjælp, mens de i andre ikke får hjælp. Samtidig er der betydelige forskelle i befolkningen mellem kommunerne, så de 85-årige i nogle kommuner er forholdsvis bedre fungerende end de 85-årige i andre kommuner. Det betyder, at det kan være vanskeligt at fortolke betydningen af de kommunale variabler uanset fortegn og størrelse, da der ofte vil være modsat rettede effekter.

Estimationen af sandsynligheden for at modtage hjemmehjælp har derfor følgende udseende:

$$\text{Tildeling} = \alpha_0 + \alpha_1 * \text{alder} + \alpha_2 * \text{alder}^2 + \alpha_3 * \text{civil} + \alpha_4 * \text{køn} + \alpha_5 * \text{Tid-til-død} + \alpha_6 * \text{afstand-til-barn} + \alpha_7 * \text{boligtype} + \alpha_8 * \text{kommune} + \varepsilon$$

Populationen består af alle over 65 år d. 1. januar 2012, hvilket åbner for personer, som er døde de næste fem år. Der indgår 79 kommuner i materialet, hvilket svarer til 84 pct. af befolkningen i aldersgruppen. Det er valgt at bruge dummies for såvel tid-til-død som afstand-til-barn, da tid-til-død og afstand-til-barn for nogle er ukendte eller manglende. Tilsvarende er boligtypen modelleret ved brug af tre dummies i stedet for størrelsen af bolig kombineret med ejerforhold.

Alder er modelleret som alder efter 65 år og ikke som alder siden fødsel, hvilket påvirker den funktionelle form. Det har især betydning for den kvadrerede alder, som vil stige kraftigt ved brug af alder siden fødsel. Med en anden udformning vil resultatet formodentlig være anderledes.

De estimerede koefficienter viser også, at alder er en meget afgørende parameter for, hvem der får hjemmehjælp, jf. tabel 1. Det er dog sværere at vurdere betydningen af alder som følge af dens kontinuerte form.

Tabel 1: Hovedresultater af estimation af at få hjemmehjælp, 2012

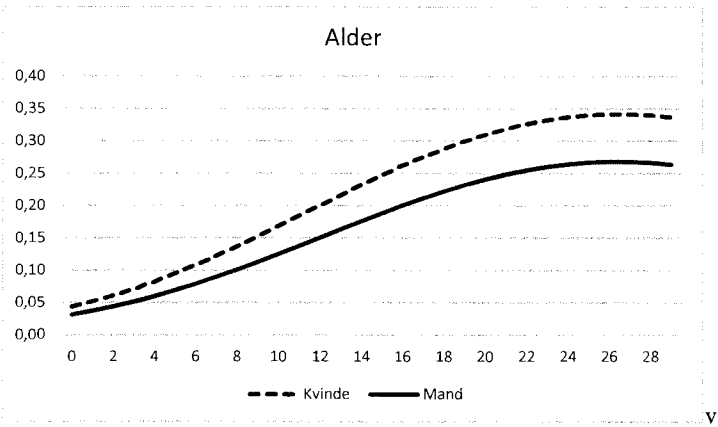
Variabel	Koefficient	Standard fejl

Konstant	-1,986	*0,025*
Alder siden 65 år	0,184	*0,002*
Alder siden 65 **2	0,0035	*0,0001*
Køn (mand)	-0,348	*0,008*
Lille lejebolig	0,342	*0,009*
Stor lejebolig	0,420	*0,010*
Lille ejerbolig	0,202	*0,017*
Gift	-0,939	*0,008*
Ugift par	-0,807	*0,023*
Død i 2013	0,187	*0,017*
Død i 2014	0,063	*0,017*
Død i 2015	-0,019	*0,017*
Død i 2016	-0,068	*0,017*
Død i 2017	-0,160	*0,018*
Død efter 2017	-1,106	*0,013*
Obs=810.000		

Note: Koefficienter for tid-til-død og kommune er ikke medtaget. Referencepersonen er en enlig kvinder bosiddende i en stor ejerbolig i København kommune, som døde i 2012.
Kilde: Egne beregninger på baggrund af registerdata.

Alder har en stor betydning for, om en person forventes at få hjemmehjælp, jf. figur 4. Uden at inddrage andre forhold er sandsynligheden for at en kvinde på 65 år i en stor ejerbolig i København kommune modtager hjemmehjælp knap 5 pct., mens den er 34 pct. for at en kvinder på 89 år i samme situation får hjemmehjælp.

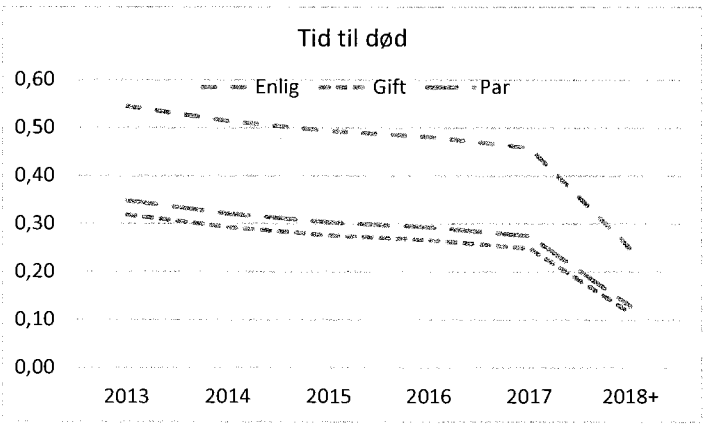
Figur 4: Betydningen af alder for visitation af hjemmehjælp, 2012



Note: Referencepersonen er en enlig person bosiddende i en stor ejerbolig i København kommune, som levede ved udgangen af 2017.
Kilde: Egne beregninger på baggrund af registerdata.

Beregningerne viser også, at civilstanden spiller en betydelig rolle for sandsynligheden for, at en ældre person modtager hjemmehjælp, og enlige er næsten dobbelt så sandsynlige modtagere af hjemmehjælp som en ægtefælle eller en partner, jf. figur 5. Her er sandsynligheden for en enlig kvinde illustreret sammen med sandsynligheden for en kvindelig ægtefælle og en kvindelig partner, og samtidig er betydningen af tid-til-død vist. Tid-til-død har betydning, men forskellen i sandsynlighed mellem personer, som døde i 2013 og efter 2017 er ikke så stor som forskellen på en 61 årig og en 84-årig.

Figur 5: Sandsynlighed for hjemmehjælp, civilstand og tid-til-død



Note: Referencepersonen er en kvinde bosiddende i en stor ejerbolig i Københavns kommune.
Kilde: Egne beregninger på baggrund af registerdata.

Modellen er langt fra at være perfekt til at beregne, hvem som modtager hjemmehjælp, idet den forudsiger knap en ud af tre positive observationer korrekt. Indikatorerne for funktionsevne og helbred er langt fra at være informative. Det letteste er uden tvivl at beregne, hvem som ikke får hjemmehjælp, hvor knap 95 pct. er beregnet korrekt. Det gælder især de yngste i populationen, hvor kun en lille andel af personerne under 75 år modtager hjemmehjælp.

Afstanden til det nærmeste barn blev indføjet som en variabel, der kunne forklare, at nogle ældre med nært boende børn ikke fik hjemmehjælp, selvom pensionister med samme karakteristika uden børn i nærheden får hjemmehjælp. Hovedparten af dummies for afstanden til nærmeste barn er ikke signifikante, men variablen for ingen børn i Danmark viser en svagt højere sandsynlighed for at få hjemmehjælp set i forhold til sandsynligheden for at få hjemmehjælp, hvis et barn bor i samme sogn som personen. Beregnet for en enlig kvinde på 80 år i København er forskellen på 2 pct.,

men for en enlig kvinde i København med et barn i et andet sogn under 10 km fra eget sogn sandsynligheden for at få hjemmehjælp 4 pct. mindre. De resultater viser ikke de forventede resultater, og konklusionen kan være, at børn ikke spiller samme rolle for de ældres tilværelse som tidligere. Til gengæld kan det offentlige have fået en større rolle for pleje og hjælp til de ældre. Det svarer til standardbeskrivelsen af den universelle velfærdsmodel.

De geografiske variabler er typisk signifikante ved et 5 pct. niveau, og niveauet svinger en del. En enlig kvinde på 80 år med død i 2013 har således ca. 40 pct. sandsynlighed for hjemmehjælp i en kommune, mens den er over 65 pct. i en anden kommune. Hovedparten af kommunerne i beregningen har en sandsynlighed mellem 0,50 og 0,55 for en kvinde med den profil.

3. Hvor meget hjemmehjælp får de

Andelen af borgere, som modtager hjemmehjælp er stigende med alderen, men det gennemsnitlige antal timer pr. modtager stiger først for modtagere over 90 år. Det gør det umiddelbart mere besværligt at estimere antallet af timer, som de modtager i hjemmehjælp.

I estimationen er medtaget alle borgere med et positivt antal timer i hjemmehjælp og de medtagne variabler er de samme. Variablene for afstand til nærmeste barn er igen udeladt, hvor halvdelen ikke er signifikante ved et 5 pct. niveau. Ligeledes er kommunevariablerne udeladt, fordi der er så mange. Det kan dog bemærkes, at nogle er meget positive, mens relativt få er negative.

Estimationen af omfanget af tildelt hjemmehjælp viser, at det med de nuværende forklarende variabler, så forklarer modellen kun en meget lille del af variationen, idet $R^2 = 0,06$. Dette niveau betyder, at det er en lille del af variationen i hjemmehjælp, som forklares af de medtagne variabler.

En af de betydelige variabler er alder og alder kvadreret. Disse to variabler er meget signifikante og har stor indflydelse på det endelige resultat. Derimod er variablen for køn ikke signifikant, hvilket kan være lidt overraskende i lyset af, at middellevetiden for mænd er kortere end for kvinder, men mænd får til trods for dette hjemmehjælp i det samme omfang som kvinder.

Boligens ejerforhold og størrelse har tilsyneladende en interessant sammenhæng med omfanget af hjemmehjælp, idet pensionister i mindre lejebolig modtager mere hjemmehjælp end i de øvrige boligtyper. En del ældre sælger deres ejerbolig og flytter i ældrevenlige plejeboliger, hvor det er lettere at få hjemmehjælp, hvilket måske kan forklare det større omfang af hjemmehjælp til personer i mindre lejeboliger.

Tabel 2: Koefficienter fra OLS estimation, 2012

Variabel	Koefficient	Standard fejl

Konstant	158,94		5,42
Alder siden 65 år	-8,82	**	0,35
Alder siden 65 **2	0,328	**	0,011
Køn (mand)	-1,05		1,71
Lille lejebolig	15,83	**	1,94
Stor lejebolig	-8,56	**	2,20
Lille ejerbolig	-12,64	**	3,87
Gift	8,38	**	1,98
Ugift par	-13,58	**	5,84
Død i 2013	107,27	**	3,19
Død i 2014	72,78	**	3,31
Død i 2015	47,61	**	3,37
Død i 2016	30,35	**	3,44
Død i 2017	8,95	**	3,54
Død efter 2017	-45,60	**	2,62
Obs=139.000	R^2= 0,06		

**: Signifikant ved et niveau på 5 pct. Referencepersonen er en enlig kvinde med bopæl i en større ejerbolig, som er død i løbet af 2012.
Kilde: Egne beregninger på baggrund af registerdata.

Gifte personer modtager mere hjemmehjælp end enlige, og det laveste omfang af hjemmehjælp tildeles samboende par, som ikke er gift. Her er det interessant, at mens gifte langt sjældnere modtager hjemmehjælp i henhold til logit beregningen, så modtager de gifte med hjemmehjælp et større omfang af det, end det er tilfældet for de enlige. En fortolkning kunne være, at gifte kan klare mindre problemer selv, mens de til gengæld får behov for mere hjælp, når de ikke kan klare det selv. Hvorfor de ugifte kan klare sig med mindre hjemmehjælp end enlige, og de enlige oftere får hjemmehjælp end de ugifte par, kan virke overraskende. Der er i analysen ikke skelnet mellem personer, som har været samboende i en årrække, og personer som kun har boet sammen i en kortere årrække.

Omfanget af hjemmehjælp til personer, som er døde de følgende år, viser det forventede mønster, idet personerne, som dør i 2013, får mere hjemmehjælp, end dem som dør de efterfølgende år. Omfanget af hjemmehjælp er aftagende, jo længere at borgerne lever, hvilket er det forventede resultat.

Variablerne for afstanden til nærmeste barn er som hovedregel ikke signifikante, hvilket betyder, at omfanget af hjemmehjælp til personer med børn i samme sogn og personer med børn længere væk ikke er signifikant forskellige. Undtagelserne er personer med børn i en sogneafstand på 30-40 km, 50-75 km eller 100-150 km, som får mindre hjemmehjælp. Det kan ikke umiddelbart forklares. Samlet set indikerer estimationen, at børn ikke har nogen betydning for tildelingen af hjemmehjælp.

De kommunale variationer er ikke de samme som ved tildelingen af hjemmehjælp. Det betyder, at de kommuner, som forholdsvist sjældent tildeler hjemmehjælp, de tildeler ikke signifikant mindre end referencekommunen. Den kommune, som oftest tildelte

hjemmehjælp, giver også mere end referencekommunen, men den ligger ikke i toppen på dette område.

4. Sammenfatning

Tildeling af hjemmehjælp til ældre borgere sker ud fra de ældres behov, men tildelingskriterierne fastsættes i de enkelte kommuner. Da den demografiske sammensætning i kommunerne er vidt forskellige, stiller det meget forskellige krav til kommunernes økonomi. Nogle kommuner som Samsø, Lolland og Hørsholm har mere end 3 pct. af befolkningen over 85 år, mens andre som Skanderborg, Solrød og Egedal har mindre end 1,5 pct. over 85 år, og denne forskel bliver ikke mindre de kommende år. Det betyder, at udgifterne til hjemmehjælp kan være vidt forskellige for kommunerne.

Indbyggernes helbred og økonomiske forhold kan også variere betydeligt, så funktionsevnen og de økonomiske muligheder kan variere betydeligt mellem eksempelvis Langeland og Gentofte. Det kan have betydning for behovet for hjemmehjælp.

Generelt er alder dog den vigtigste forklarende faktor for, om borgerne modtager hjemmehjælp, og i hvilket omfang at de modtager hjemmehjælp. Dødstidspunkt har nogen betydning for tildelingen af hjemmehjælp, mens den klart influerer på omfanget af hjemmehjælp. Civilstand har betydning for, om borgeren får tildelt hjemmehjælp, men i mindre grad omfanget af hjemmehjælp.

Afstanden til det nærmeste barn blev introduceret som en indikator for børnene bidrag til at hjælpe de ældre, men enden er variablen dårlig, eller så er børnene ikke en hjælp, som influerer på tildelingen af hjemmehjælp. Det tyder på, at den universale velfærdsmodel ikke tildeler børnene et ansvar for forældrenes velfærd.

Modellen kunne udbygges med en mere direkte beskrivelse af borgernes helbred, hvor en af de tilgængelige variabler er antallet af besøg hos alment praktiserende læge eller antallet af indlæggelsesdage. Begge variabler kunne findes for året før.

5. Litteratur

Dalgaard, C-J. og H. Strulik (2014): Optimal Aging and Death: Understanding the Preston Curve.

How far does the peer effect on smoking behavior reach?

Jørgen T. Lauridsen, Department of Business and Economics, University of Southern Denmark, E-mail jtl@sam.sdu.dk.

Abstract

Peer effects in forming smoking habits are well known from literature. If living in a low smoking environment, an individual will be less tempted to smoke, and reversely if living in a high smoking environment. Based on municipality level data for 98 Danish municipalities observed in three waves in 2010, 2013 and 2017, the present study investigates whether such peer effects can be found between municipalities in the form of an endogenous spatial spillover. By using a spatial quantile regression approach, it is investigated whether peer effects are present in extreme environments (i.e. environments with low or high smoking propensities) as well as environments with a median smoking propensity. A peer effect in the form of a significantly positive spatial spillover is found for median environments. For extreme environments with low or high smoking propensities, a positive, but not significant, peer effect is found. The results indicate that the peer effect predominantly exists in environments characterized by a median smoking behavior, and that it is less relevant for environments with high or low smoking propensities.

Keywords:

Smoking; spatial spillover; peer effects; quantile regression

JEL classifications:

C13, C21, C23, I12, I14

1. Introduction

Peer effects in forming smoking habits are well known, especially among adolescents, but also among the general population (ref). Thus, if living in a low smoking environment, an individual will be less tempted to smoke, and reversely if living in a high smoking environment. This readily translates into a spatial peer effect setup, which can be embedded in an endogenous spatial spillover setup, i.e., smoking prevalence will be higher in an area if smoking prevalence is high in surrounding areas.

However, while spatial patterns of smoking prevalence are analyzed in literature (ref), no studies, to my knowledge, has focused on spatial spillover of smoking behavior. The present study aims at suggesting such a framework together with an empirical analysis for the case of Denmark. Specifically, based on municipality level data for 98 Danish municipalities observed in three waves in 2010, 2013 and 2017, the present

study investigates whether spatial peer effects can be found between municipalities in the form of an endogenous spatial spillover.

By using a spatial quantile regression approach, it is investigated whether peer effects are present in extreme environments (i.e. environments with low or high smoking propensities) as well as environments with a median smoking propensity. The study also controls for a variety of explanatory variables catching up on health status, health behavior and sociodemographic characteristics. Based on literature and availability of data, health was measured by variables describing physical and mental illness (ref). Similarly, health behavior involved population propensities for physical exercise, unhealthy cost, alcohol and obesity. Finally, sociodemographic characteristics involved age distribution, urbanization degree, education, social burdens, income and whether the municipal is a peripheral one.

The structure of the paper is as follows. The following section 2 outlines the econometric model, i.e., a spatial quantile regression model. Next, data are briefly summarized and described in Section 3, while Section 4 presents results. Specifically, it is demonstrated that a spatial peer effect in the form of a significantly positive spatial spillover is found for median environments, while for extreme environments with low or high smoking propensities, a positive, but not significant, peer effect is found. Finally, Section 5 rounds off with concluding remarks and recommendations.

2. Methods

The study applies quantile regression which estimates functional relations between variables for selected portions of probability distribution. An ordinary least squares (OLS) regression is based on the mean of the conditional distribution of the dependent variable. When exogenous variables influence the parameters of the conditional distribution of the dependent variable other than the mean, OLS is not sufficiently robust (see Koenker and Bassett, 1978). Unlike OLS, quantile regression models include a full characterization of the conditional distribution of the dependent variable.

The most important properties of quantile regressions are as follows: they are strong against the outlier or tail value of the dependent variable and offer more effective estimates when the error term is not normally distributed and enables analysis of any point of distributions. Quantiles are generally illustrated as τ and $0 < \tau < 1$.

Quantile regression is not yet applicable to estimation of a spatial error model. Thus, several researchers have proposed following variants of the spatial autoregressive (AR) model

$$Y = \rho(\tau)WY + X\beta(\tau) + u \qquad (1)$$

where W is an $(N \times N)$ spatial weight matrix, while ρ indicates the degree of spatial autocorrelation and is a parameter to be estimated. τ indicates the corresponding

quantile, and $\rho(\tau)$ and $\beta(\tau)$ are to be estimated. Y is $(N\times1)$ a dependent variable, X is $(N\times k)$ an explanatory variable, and u is error term.

Kim and Muller's (2004) estimation method is one of the methods used for estimating the model on (1). The estimation method requires two stages. In the first stage, the spatially lagged endogenous variable WY is regressed on the spatially lagged exogenous variables WX as well as X for a quantile. The predicted value of WY is substituted for WY in the spatial-lag model. Thus, the correlation between the spatially lagged endogenous variable and the error term is eliminated. The second stage is a quantile regression of Y on X and the predicted values of WY. This two-stage procedure is repeated for other quantiles. They use bootstrap procedures to construct standard error estimates.

3. Data

Table 1 provides an overview of the data used for the study.

Table 1. Description of data.

Variable	Description	Mean	SD
DailySmoker [1]	Percentage reporting being daily smokers	18.71	3.58
year13	Indicator for year 2013 (reference year 2010)	0.33	0.47
year17	Indicator for year 2017 (reference year 2010)	0.33	0.47
PhysAct [1]	% reporting being physically active in leisure time	27.19	3.30
Alcohol [1]	% reporting drinking too much (21/14 glasses weekly for M/F)	8.47	2.33
UheFood [1]	% reporting easting unhealthy food daily	15.08	3.77
Obese [1]	% reporting being obese (BMI>30)	51.22	5.66
SAH [1]	% reporting good / very good Self Assessed Health	83.85	2.80
Stress [1]	% reporting being stressed daily	13.29	2.76
IllPhysHealth [1]	% reporting ill physical health	11.39	2.45
IllMentalHealth[1]	% reporting ill mental health	10.68	2.18
UnwantedAlone[1]	% reporting being unwanted alone daily	5.62	1.09
P25_64 [3]	% population between 25 and 64	51.19	1.88
P_65 [3]	% population 65 and above	19.69	3.87
Urban [3]	% population living in urban area	83.22	12.77
SocHous [3]	% of population living in social housing	18.05	12.01
Educ [3]	% of population with higher education	24.48	8.76
Unemp [3]	% of population unemployed	4.01	1.12
NonWest [3]	% of population being from non-Western country	3.10	1.92
SiCrime [3]	Simple crimes reported per 1,000 inhabitants	48.48	19.82
ViCrime [3]	Violence crimes reported per 1,000 inhabitants	1.57	0.62
SocBen [3]	% population receiving social benefits	4.12	1.25
TaxBase [3]	Tax deductible income per inhabitant (100,000 DKK, 2010 level)	1.72	0.34
Peripheral [2]	Indicator for peripheral municipality	0.15	0.36

Source: [1] The Danish Health Profile (www.danskernessundhed.dk), [2] The Danish Ministry of Taxation (www.skm.dk), and [3] the Key Figure Base (www.im.dk)

Data are collected for 98 municipalities for the years 2010, 2013 and 2017 from three different sources, The Danish Health Profile, The Danish Ministry of Taxation, and the

Key Figure Base at The Ministry of the Interior. The outcome variable is percentage reporting being daily smokers. According to theory, literature and availability, a variety of explanatory variables affecting smoking propensity were included. As an indication of change over time, dummies for 2013 and 2017 are defined. Health behaviour includes physical activity, alcohol, unhealthy food habits and obesity, while health and stress effects are captured by Self-Assessed Health (SAH), stress, ill physical health and ill mental health. Social relations are represented by one variable (being unwanted alone), and age structure by percentages of 25-64 and 65+ year. Finally, presence or absence of social stress is captured by a variety of variables (urbanisation, social housing, education, unemployment, non-Western inhabitants, criminal activity, social benefit receivers, income level and peripheral area).

4. Results

Table 2 shows the results from the spatial quantile regressions.

Table 2. Spatial quantile regressions for Daily Smoking

Quantile	0.10		0.25		0.50		0.75		0.90	
	Est	T	Est	T	Est	T	Est	T	Est	T
Intercept	52.5***	2.56	19.8	1.14	10.3	0.81	22.5*	1.86	6.19	0.34
W*DailySmoker	0.12	1.20	0.13	1.53	0.18**	2.24	0.01	0.02	0.17	1.28
year13	-4.11***	-5.98	-3.85***	-6.30	-3.39***	-6.56	-4.07***	-6.86	-3.83***	-4.87
year17	-5.17***	-4.27	-5.53***	-5.17	-4.89***	-5.37	-5.63***	-5.69	-5.17***	-3.76
PhysAct	-0.18	-1.61	-0.09	-1.40	-0.06	-1.03	-0.12*	-1.89	-0.16*	-1.84
Alcohol	0.08	0.71	0.17*	1.88	0.22***	2.77	0.21***	2.39	0.15	1.34
UhealthyFood	0.13**	2.18	0.10*	1.74	0.12**	2.35	0.06	1.23	0.07	0.84
Obese	-0.05	-0.78	-0.01	-0.00	-0.02	-0.58	-0.09*	-1.80	-0.01	-0.08
SAH	-0.37**	-2.42	-0.23*	-1.91	-0.09	-0.90	-0.15	-1.31	-0.03	-0.23
Stress	0.22*	1.88	0.24***	2.57	0.22***	3.04	0.32***	3.65	0.14	1.09
IllPhysHealth	-0.07	-0.40	-0.06	-0.52	0.16	1.33	0.04	0.40	0.01	0.08
IllMentalHealth	0.01	0.05	0.04	0.30	-0.01	-0.15	-0.02	-0.25	0.05	0.28
UnwantedAlone	-0.08	-0.44	0.09	0.70	0.07	0.61	0.04	0.33	0.20	1.06
P25_64	0.08	0.38	0.27*	1.93	0.25**	2.49	0.39***	3.21	0.30*	1.87
P_65	0.04	0.35	0.17*	1.72	0.14	1.63	0.20**	2.38	0.21*	1.88
Urban	-0.03	-1.16	0.01	0.09	-0.02	-1.47	-0.03**	-1.97	-0.02	-0.95
SocHous	-0.04	-1.45	-0.02	-1.09	0.01	0.06	0.01	0.58	0.01	0.27
Educ	-0.06	-1.27	-0.10***	-2.70	-0.09***	-3.44	-0.11**	-2.35	-0.02	-0.37
Unemp	-0.25	-1.29	-0.25	-1.44	-0.35**	-2.49	-0.63***	-4.08	-0.37	-1.60
NonWest	0.48	0.28	0.10	0.08	0.09	0.10	0.05	0.06	-0.14	-0.07
SimpleCrime	0.01	0.22	-0.01	-1.33	-0.01	-0.34	-0.01	-1.42	-0.01	-0.32
ViolentCrime	0.02	0.07	-0.11	-0.47	-0.11	-0.60	-0.14	-0.78	-0.49	-1.32
SocBenefit	0.58***	2.53	0.64***	4.18	0.48***	3.82	0.51***	3.41	0.80***	2.97
TaxBase	-0.83	-0.65	-0.53	-0.44	-0.86	-0.98	-2.17**	-2.01	-1.71*	-1.74
Peripheral	-0.05	-0.13	-0.03	-0.09	0.20	0.69	0.55	1.55	0.30	0.52

Note. Significance indicated by ***, ** and * for 1, 5 and 10 percent significance respectively

Turning to the main question, i.e. as to whether there is a spatial peer effect, the spatial lag of DailySmoker is positive throughout thus indicating a spatial peer effect. However, it is only significant for the median regression, thus indicating that the peer

effect is less in play for extreme environments with high or low proportions of smokers.

Besides, smoking behavior is to some extent affected by different variables. First, there seems to be a general reduction in smoking behavior throughout over time, as the level is lower for 2013 and 2017 for all percentiles. Next, lifestyle seems to correlate to some extent to smoking behavior. Phsysical activity is negatively related to smoking, but only significantly so for the high smoking percentiles. Alcohol consumption is positively related, but predominantly for the median and quartile smoking percentiles. Likewise, unhealthy food consumption correlates positively with smoking, but predominantly for the median and lower smoking percentiles. Physical and mental health seems to some extent related to smoking, as the proportion reporting good SAH is negatively related to smoking, however only significantly so for the lower smoking percentiles. The proportion being stressed seems significantly positively related to smoking, unless for the extremely high smoking percentile. On the other hand, the effect of loneliness is not significant for any smoking percentile. Age seems to play a role, as proportions of 25-64 and 65+ year old is positively related to smoking behavior, however not significantly so for the low smoking percentile. Furthermore, high education seems to be negatively related to smoking, but not significantly so for the upper or lower extreme smoking percentiles. Turning to indicators of social stress, it is a bit surprising to see that smoking is negatively related to unemployment. On the other hand, high proportions of social benefit receivers is strongly related to smoking propensity for all smoking percentiles. Finally, high income levels appear negatively related to smoking, but only significantly so for the higher smoking percentiles.

5. Conclusion

The innovative contribution of the present study is twofold. First, the study adds by suggesting a spatial peer effect in smoking behavior, operationalized as an endogenous spatial spillover effect of smoking prevalence. Second, it is demonstrated by combining with a quantile regression that the peer effect predominantly exists in environments characterized by a median smoking behavior, and that it is less relevant for environments with high or low smoking propensities.

As a secondary outcome, different variables were shown to be significantly related to smoking behavior. Only few of these (years 2013 and 2017 levels and proportion of social benefit receivers) related significantly to all smoking percentiles. Some (unhealthy food consumption, self-assessed good health, and stress) related predominantly to the lower smoking percentiles, while other (physical activity, alcohol, tax base, and proportions of 25-64 and 65+ years) related significantly to upper smoking percentiles, and only two (education and unemployment) merely related significantly to median and quartile smoking percentile. The signs of these significant effects were as expected, although with an exception for unemployment,

which, surprisingly, turned out to be significantly negatively related to the median and 75 percent smoking percentiles.

References

Koenker, R. and Bassett, G. (1978). "Regression quantiles". Econometrica, 46, 33-50.

Kim, T.H. and Muller C. (2004). "Two-Stage Quantile Regression when the First Stage is Based on Quantile Regression". Econometrics Journal, 7, 218-231.

Latente klasser af DSM-5 AUD blandt danske ældre i alkoholbehandling

Anna Mejdal, SDU

Baggrund: En voksende del af ældre har et stort forbrug af alkohol. Studier på "Baby-boomer"-generationen, født efter anden verdenskrig, indikerer at denne generations alkoholforbrug, også oppe i alderen, fortsætter på et højere niveau end tidligere generationers. Kun ganske få undersøgelser belyser ældres alkoholvaner i Danmark.

Formål: Vi udviklede latente klasser af DSM-5 Alcohol Use Disorder (AUD) kriterierne (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition) blandt danske voksne over 60 år, som opsøgte alkoholbehandling, og undersøgte om forskellige baggrundsfaktorer og behandlingsresultater er associerede med klassetilhørighed.

Metode: Vi brugte danske deltagere fra Elderly undersøgelsen (>=60 år, n=341). Latent-klasse-analyse blev benyttet på de 11 kriterier fra DSM-5, som bruges til diagnosticering af AUD. Derefter undersøgte vi sammenhængen mellem klassemedlemskab og sociodemografiske variable, drikkemønster samt WHO Quality of Life vha. latent-klasse 3-step-metoden. Afslutningsvis undersøgte vi også, om klassemedlemskab har indflydelse på behandlingsresultat.

Indledning

The American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM) anvendes i kliniske sammenhænge til at belyse et problematisk alkoholforbrug. I den fjerde version (DSM-IV) adskilte man problematisk alkohol forbrug i to separate kategorier: misbrug og afhængighed, hvor tre ud af 7 afhængighedskriterier resulterede i en afhængighedsdiagnose, og et ud af fire misbrugskriterier resulterede i en misbrugsdiagnose, hvis afhængighedsdiagnosen ikke kunne opfyldes (American Psychiatric Association, 1994).

I 2013 blev den femte reviderede version af Diagnostic and Statistical Manual of Mental Disorders (DSM-5) (American Psychiatric Association, 2013) udgivet. På baggrund af forskellige studier af DSM-IV og anbefalinger, blev der på alkoholområdet ændret i diagnosekategorierne og diagnosemetoden (Hasin et al., 2013). Under DSM-5 blev de to diagnoser kombineret til én dimensional diagnose "Alcohol Use Disorder" AUD. Derudover blev et misbrugskriterie, som omhandler "juridiske problemer", fjernet og der blev tilføjet et kriterie om trang. Der blev etableret en tri-kategoriseret skala, hvor den diagnostiske tærskel er mindst to ud af 11 kriterier. Således defineres to til tre AUD-symptomer som mild AUD, 4-5 symptomer definerer moderat AUD, og 6 eller flere symptomer defineres som alvorlig AUD.

Identifikation af forskellige klasser af AUD ved hjælp af latent klasseanalyse

I de senere år er latent klasseanalyse (LCA) (McCutcheon, 1987) blevet brugt til at undersøge, om forskellige klasser for afhængigheds- /misbrugskategorier faktisk kan identificeres og danne grundlag for en mere meningsfuld kategorisering end udelukkende at kigge på antallet af opfyldte kriterier. Flere studier har undersøgt DSM-IV kriterierne vha. LCA [4-8], men fandt ikke støtte til at opdele disse i to forskellige afhængigheds- / misbrugskategorier.

Efter den reviderede version af DSM har yderligere fire studier udforsket heterogeniteten af DSM-5-kriterier vha. LCA i forskellige populationer (Casey, Adamson, & Stringer, 2013; Castaldelli-Maia et al., 2014; Rinker & Neighbors, 2015; Swift et al., 2016). Fælles for alle fire studier er en klassestruktur, der primært repræsenterer en stigende grad af opfyldte DSM-5 AUD-kriterier, mere end bestemte mønstre af opfyldte kriterier. Indtil videre har ingen undersøgelser undersøgt mønstre af DSM 5-kriterier ved hjælp af LCA i en kliniske stikprøve.

Aldersspecifikke kliniske populationer

Den kliniske population af personer med et problematisk alkoholbrug er en varieret gruppe, og forskning vise,r at andelen af ældre i behandling er voksende (Nielsen, Nielsen, Lolk, & Andersen, 2010). De fleste lande i den vestlige verden har stigende forventet levealder og en aldrende befolkning, og alkoholforbruget er højt blandt ældre sammenlignet med tidligere generationer (Connor, Haber, & Hall, 2016). Derfor forventes også en stigning af ældre i behandling i takt med at antallet af ældre voksne stiger (DiBartolo & Jarosinski, 2017; Han, Moore, Sherman, Keyes, & Palamar, 2016).

Ældre er mere udsatte end yngre, når det kommer til de fysiologiske virkninger af at drikke alkohol (Anderson, Scafato, & Galluzzo, 2012; Barry & Blow, 2016). Forskning har imidlertid også vist, at alkoholmisbrugsproblematikker blandt ældre, der søger behandling, ofte er mindre alvorlig end hos midaldrende og yngre individer, og at behandling af ældre med AUD har højere succesrate (Lemke & Moos, 2003; Oslin, Pettinati, & Volpicelli, 2002; Satre, Mertens, Arean, & Weisner, 2004; Wieben, Nielsen, Nielsen, & Andersen, 2018).

Det kan dog antages, at den kliniske gruppe af ældre voksne er endnu mere heterogen end den kliniske gruppe af midaldrende individer (Fingerhood, 2000; Kuerbis, Moore, Sacco, & Zanjani, 2016), og at den både omfatter individer med alkoholproblematikker, der er opstået sent i livet, men også individer med alkoholproblematikker, der har fulgt dem igennem meget af deres voksne liv (Bakhshi & While, 2015; Emiliussen, Nielsen, & Andersen, 2017). Det er således sandsynligt, at forskellige grupperinger kan identificeres i den kliniske gruppe af ældre i behandling, og denne viden kan være nyttig i tilrettelægning af behandling. Ingen studier har endnu undersøgt latente klasser af DSM-5 AUD i en klinisk stikprøve af ældre og relateret klassemedlemsskab til effekten af AUD-behandlingen.

Metode

Latent klasseanalyse (Latent Class Analysis LCA) (McCutcheon, 1987), en statistisk metode, er en speciel form for Mixture Models, som sigter mod at afdække ikke-observeret heterogenitet i en befolkning og finde meningsfulde grupperinger af individer, som minder om hinanden i deres respons på udvalgte variable.

I sin mest simple form grupperer LCA individer til subgrupper, som har lignende svar på et udvalgt sæt af observerede kategoriske variabler, og defineres på følgende måde.

Antag at X er en kategorisk latent variabel som antager T værdier (klasser), t er en bestemt latent klasse, $t = 1, \dots, T$. Antag at der er K kategoriske variable, og lad Y_k er den k'te variabel, k=1,2,...K. Lad $\boldsymbol{Y}$ være en vektor som indeholder udfaldsrummet af de K indikator variabler og lad $\boldsymbol{y}$ være et konkret udfald af $\boldsymbol{Y}$. En latent klasse model for sandsynligheden for at observere et bestemt $\boldsymbol{y}$ kan defineres på følgende måde:

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{t} P(X = t) P(\boldsymbol{Y} = \boldsymbol{y} | X = t) \qquad 1$$

Hvor $P(X = t)$ er sandsynligheden for at tilhøre klasse t og $P(\boldsymbol{Y} = \boldsymbol{y} | X = t)$ er sandsynligheden for udfaldet $\boldsymbol{y}$ givet at klassen er t. Dvs. at sandsynligheden for at observere $\boldsymbol{y}$ er et vægtet gennemsnit af de klassespecifikke sandsynligheder for $\boldsymbol{y}$.

I klassisk LCA antages lokal uafhængighed, dvs. at de K variable er indbyrdes uafhængige inden for hver klasse t. Under den antagelse kan ovenstående ligning 1 omskrives til:

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \sum_t P(X = t) \prod_k P(Y_k = y_k | X = t) \qquad 2$$

Parametrene af interesse er her klassestørrelserne $P(X = t)$ og de klassespecifikke responssandsynligheder på variablerne, $P(Y_k = y_k | X = t)$. Ud fra LCA modellen fås også de estimerede klasse sandsynligheder for hvert individ, her kaldet klassifikationssandsynligheder (posterior probabilities). Hvert individ kan have en delt klassetilhørighed, og klassifikationssandsynligheder med værdier større end nul for de fleste eller alle klasser.

Denne simple model kan udvides på flere forskellige måder, både til at kunne håndtere lokal afhængighed mellem de K variable, til at kunne håndtere både kontinuerte variabler $\boldsymbol{Y}$ eller blandede typer og også yderligere kovariater.

I de fleste softwarepakker estimeres LCA-modelparametrene vha. Expectation Maximization (EM) algoritmen for at bestemme Maximum Likelihood Estimater (MLE). MLE har den fordel, at under antagelse af at manglende svar ikke afhænger af selve svarene (Missing at Random MAR/ Missing Completely at Random MCAR), giver EM-algoritmen konsistente, asymptotisk effektive og asymptotisk normale estimater (Lanza, Collins, Lemmon, & Schafer, 2007).

Det er desværre ikke entydigt, hvordan man bedst bestemmer antallet af klasser (eller uobserverede subgrupper) i en LCA model. Flere statistiske informationskriterier anbefales i litteraturen, herunder Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) og flere justerede varianter af disse. Dernæst er der også likelihood-baserede tests: Vuong-Lo-Mendell-Rubin Adjusted Likelihood Ratio Test (VLMR-LRT) og Bootstrapped Likelihood Ratio Test (BLRT), begge evaluerer, om en yderligere klasse medfører en statistisk signifikant bedring i modellens fit (Nylund-Gibson & Choi, 2018; Nylund, Asparouhov, & Muthén, 2007). Udover modellens fit kan man også vurdere modellens anvendelighed, dvs. om de fundne klasser er meningsfulde, og om klasserne har en fornuftig størrelse (mere end 5-8% af stikprøven og en fornuftig absolut størrelse).

Efter en løsning er valgt, er der flere måder hvorpå dens kvalitet kan vurderes. Et populært mål er den relative entropi, defineret ved

$$E = 1 - \frac{\sum_{i=1}^{N} \sum_{j=1}^{T} \alpha_{ij} \ln \alpha_{ij}}{N \log T}$$

N er her antallet af observationer, og α_{it} er klassifikationssandsynligheden af det i'te individ til den j'te klasse. Den relative entropi er et mål for klassifikations-usikkerheden, som kan antage værdier mellem nul og et og bruges til at vurdere, om modellen generelt er god til at klassificere individer til klasser, værdier >0.8 vurderes her som "gode". Denne værdi kan sammenholdes med de gennemsnitlige posteriore klassifikationssandsynligheder per klasse (Average Posterior Probabilities AvePP), et mål for hvor godt modellen klassificerer individer til deres mest sandsynlige klasse. Det beregnes per klasse ud fra klassifikations-sandsynlighederne af de individer, der har den pågældende klasse som den mest sandsynlige, her vurderer litteraturen værdier >0.7 som værende "gode". Til sidst er de klassespecifikke respons-sandsynligheder i sig selv også et mål for kvalitet, da de giver et billede af hvor homogene klasserne er, og hvor ens individer er m.h.t. de valgte variabler (Nylund-Gibson & Choi, 2018; Nylund et al., 2007).

Efter valg af model er det ofte relevant at undersøge hvordan klasserne er relateret til kovariater og distale udfald. Det er, som nævnt foroven, muligt at tilføje kovariater direkte til modellen, men anbefales ikke (Nylund-Gibson & Masyn, 2016) , da det har vist sig at kunne lede til minder optimale løsninger.

Den mest simple metode er at kategorisere individer til deres mest sandsynlige klasse og derefter anvende denne kategorisering som almindelig kategorisk variabel i videre analyser, men dette har vist sig at være en konservativ metode, som undervurderer forholdet mellem kovariat/udfald og klassemedlemskab (Vermunt, 2010).

På nuværende tidspunkt er "best practice" at anvende enten Three-step metoden fra (Vermunt, 2010) eller BCH metoden (Bolck, Croon, & Hagenaars, 2004) . Disse justerer på lidt forskellig måde for ovennævnte bias og anbefales over standardmetoder (Heron, Croudace, Barker, & Tilling, 2015).

Data

Data stammer fra den danske del af ELDERLY studiet, ét af RESCueH-studierne (Andersen et al., 2015), hvor studiet præsenteres i detaljer. Derudover kan en karakteristik af gruppen mht. AUD findes i (Behrendt et al., in press).

ELDERLY studiet udviklede og testede et ambulant adfærdsterapi-program for DSM-5 AUD i ældre voksne (60 år eller ældre). Undersøgelsen blev udført som et multi-center randomiseret, kontrolleret studie i seks forskellige byer (Danmark: København, Odense, Aarhus, USA: Albuquerque, Tyskland: München, Dresden). Studiet blev gennemført fra foråret 2014 til foråret 2017. I alt blev 341 patienter indskrevet i de tre danske ambulante behandlingstilbud. Patienter blev randomiseret til enten Motivational Enhancement Therapy (MET) eller MET med en add-on session baseret på Community Reinforcement Approach (CRA). En række spørgeskemaer blev anvendt, herunder Form-90, Alcohol Dependence Scale, Penn Alcohol Craving Scale, Brief Symptom Inventory og WHO Livskvalitet, indsamlet gentagne gange på forskellige tidspunkter (baseline, 4 uger, 12 uger, 26 uger, 52 uger) for hver deltager. Her anvender vi kun data fra baseline og 26 uger.

Følgende spørgeskemaer blev anvendt i dette studie:

For at bedømme DSM-5 AUD blev Mini International Neuropsychiatric Interview (M.I.N.I.) version 5.0.0 brugt, udvidet med to spørgsmål om craving i alkoholafsnittet.

De 11 kriterier var TOLERANCE, WITHDRAWAL, drikke mere end beregnet (LOSS OF CONTROL), mislykkede forsøg på at reducere sit forbrug eller stoppe med at drikke (DESIRED CONTROL), bruge meget tid på at få fat på alkohol eller have tømmermænd (TIME SPENT) opgive vigtige fritidsaktiviteter (LESS ACTIVITIES), fortsætte med at drikke på trods af fysiske eller psykiske problemer (PHYSICAL/PSYCHOLOGICAL), drikke i situationer hvor man har et ansvar (RESPONSIBLE ROLE), være påvirket i situationer der er fysisk farlige (HAZARDOUS SITUATIONS), fortsat alkohol overforbrug på trods af problemer med sine omgivelser og sociale relationer (SOCIAL / INTERPERSONAL) og CRAVING.

I M.I.N.I. blev deltagerne ligeledes spurgt om hvor gamle de var, da de første gang oplevede at have mindst to af de 11 DSM-5 AUD-kriterier.

Livskvalitet blev vurderet med WHOQOL-BREF, en forkortet 26-udgave af WHOQOL-100, som er et livskvalitetsvurderingsværktøj udviklet af WHOQOL-gruppen (Skevington, Lotfy, & O'Connell, 2004). Det indeholder 26 spørgsmål, der er grupperet i fire domæner: Domæne 1: Fysisk sundhed, Domæne 2: Psykologisk helbred, Domæne 3: Sociale relationer og Domæne 4: Omgivelser og miljø. Livskvalitet blev vurderet både ved baseline og ved 26 ugers opfølgningen. Scores kan variere mellem 4 og 20, og en højere score indikerer højere livskvalitet i det pågældende domæne.

Drikkemønster blev vurderet ved hjælp af Form 90 (Miller & Boca, 1994), som spørger ind til patientens daglige alkoholindtag. Vi har beregnet antal drikkedage og genstande pr. drikkedage over en 30-dages periode på baggrund af hhv. dag 60-31 før baseline og for de 30 dage der leder op til 26-

ugers opfølgningen. Ved baseline vurderede vi, at dag 60-31 før behandlingsstart giver et mere præcist billede af patienternes alkoholforbrug præ-behandling. Ofte vil den første kontakt til klinikken være op til 3 uger før behandlingen indledes og baseline målingen foretages, samtidig med at beslutningen om at komme i behandling og den første kontakt kan have en påvirkning på alkoholforbrugets størrelse og hyppighed.

Baseline karakteristik

Tabel 1: baseline karakteristik

	Kvinder (n=122)	Mænd (n=219)	Alle(n=341)
Demografi			
Køn (kvinder,%)	-	-	35.8
Alder, median (IQR)	65 [62;69]	64 [62;68]	64 [62;68]
Samboende (ja, %)	35.3	49.8	44.6
Uddannelse (+10 år, %)	42.6	44.0	43.5
Alkohol relateret			
Drikkedage på 30 dage, median (IQR)	24.5 [6;30]	27 [13;30]	26 [12;30]
Genstande per drikkedag, median (IQR)	7.1 [4.0;10.8]	9.3 [5.6;15.0]	8.4 [5;13.5]
Tidligere behandling (ja, %)	56.6	64.8	61.8
Alder ved første AUD, median (IQR) ***	50 [40; 59.5]	45 [30; 57]	47 [35; 58]
DSM-5 AUD Diagnose (%) **			
mild	6.1	7.5	7.0
moderat	16.7	20.6	19.2
alvorlig	77.2	71.9	73.8
Quality of Life domæne score			
Fysisk sundhed, median (IQR)	12.6 [10.9;13.7]	12.6 [11.4;13.7]	12.6 [11.4;13.7]
Psykologisk helbred, median (IQR)	12.7 [11.3;14.0]	13.3 [12.0;14.7]	13.3[12;14.7]
Sociale relationer, median (IQR)	14.7 [13.3;16.0]	13.3 [12.0,16.0]	14.7 [12;16]
Miljø og omgivelser, median (IQR)	16.5 [14.5; 18.5]	17.0 [16.0;18.5]	16.5 [15.5;18.5]

* 43 missing (26,17) pga. non-response, ** 28 missing (20,8) pga. fejlregistrering, *** 80 missing (58,22) da patienter ikke kan huske alder.

Analyse

For at vælge det rigtige antal klasser og maksimere model fit, blev 1-6 klasse-modeller beregnet. BIC, AIC, SABIC (sample size adjusted BIC) og CAIC (Consistent AIC) blev sammenlignet, se tabel 2. Derudover valgte vi at sammenligne klasser vha. BLRT-testen. Grundet stikprøvens størrelse skulle ingen klasse skulle have færre end 10% (ca. 30 personer) i den.

Tabel 2: Model fit statistik

	LL	BIC(LL)	AIC(LL)	CAIC(LL)	SABIC(LL)	df	p-value*
1-Klasse model	-1959.98	3984.11	3941.96	3995.11	3949.22	330	
2-Klasse model	-1879.18	**3892.49**	3804.36	**3915.49**	3819.53	318	<0.001
3-Klasse model	-1859.28	3922.67	3788.56	3957.67	3811.65	306	<0.001
4-Klasse model	-1845.18	3964.46	3784.36	4011.46	3815.37	294	0.07
5-Klasse model	-1834.60	4013.29	3787.21	4072.29	3826.13	282	0.32
6-Klasse model	-1826.02	4066.11	3794.05	4137.11	3840.89	270	0.49
3-Klasse model justeret**	-1844.50	3916.44	**3766.99**	3955.44	**3792.72**	302	<0.001***

*p-værdi fra BLRT-testen

** justeret med direkte effekter mellem (Loss of Control og Less activities), (Loss of Control og Craving), (Physical/Psychological og Tolerance), (Physical/Psychological og Interpersonal)

*** sammenlignet med 3-klasse model uden justeringer, BLRT p-værdi var 0.48 i sammenligning af 4-klasse-model med den justerede 3-klasse model.

Informationskriterierne BIC og CAIC favoriserede 2-klassemodellen, AIC 4-klassemodellen og SABIC 3-klassemodellen. BLRT testen fandt at 3-klasse modellen havde bedre fit end 2-klasse-modellen, men ikke et bedre fit ved 4-klasse modellen på et 5% niveau (p = 0,07).

Vi fandt, at betingelsen om lokal uafhængighed ikke var opfyldt i fire tilfælde i 3-klasse modellen og justerede derfor for: Loss of Control og Less activities; Loss of Control og Craving; Physical/Psychological og Tolerance; og sidst Physical/Psychological og Interpersonal. Denne model blev valgt som den endelige løsning.

Entropi for den endeligt justerede 3-klasse model var lav ved 0.59, hvilket indikerer at klasseseparation ikke er på et højt niveau. På den anden side varierede AvePP fra 0.80 til 0.83, hvilket tyder på, at selv om ikke alle patienters DSM-5 AUD-mønstre er perfekt repræsenteret af de tre klasser, er tildelingen til klasser relativt god.

Alle analyser blev udført i Stata 15 og LatentGold 5.1.

De tre klassers karakteristik og association til kovariater og behandlings-outcome præsenteres ved symposiet.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.). Washington, DC: American Psychiatric Association.

Andersen, K., Bogenschutz, M. P., Buhringer, G., Behrendt, S., Bilberg, R., Braun, B., . . . Nielsen, A. S. (2015). Outpatient treatment of alcohol use disorders among subjects 60+ years: design of a randomized clinical trial conducted in three countries (Elderly Study). *BMC Psychiatry, 15*, 280. doi:10.1186/s12888-015-0672-x

Anderson, P., Scafato, E., & Galluzzo, L. (2012). Alcohol and older people from a public health perspective. *Ann Ist Super Sanita, 48*(3), 232-247. doi:Doi: 10.4415/ann_12_03_04

Bakhshi, S., & While, A. E. (2015). Older people and alcohol use. *Gastrointestinal Nursing, 13*(Sup10), S27-S32. doi:10.12968/gasn.2015.13.Sup10.S27

Barry, K. L., & Blow, F. C. (2016). Drinking Over the Lifespan: Focus on Older Adults. *Alcohol Res, 38*(1), 115-120.

Behrendt, S., Braun, B., Bilberg, R., Buehringer, G., Bogenschutz, M. P., Nielsen, A. S., Mejldal, A., & Andersen, K. (in press). DSM-5 alcohol use disorder features among treatment-seeking older adults. *Sucht*.

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators. *Political Analysis, 12*(01), 3-27.

Casey, M., Adamson, G., & Stringer, M. (2013). Empirical derived AUD sub types in the US general population: a latent class analysis. *Addict Behav, 38*(11), 2782-2786. doi:10.1016/j.addbeh.2013.06.022

Castaldelli-Maia, J. M., Silveira, C. M., Siu, E. R., Wang, Y. P., Milhoranca, I. A., Alexandrino-Silva, C., . . . Martins, S. S. (2014). DSM-5 latent classes of alcohol users in a population-based sample: results from the Sao Paulo Megacity Mental Health Survey, Brazil. *Drug Alcohol Depend, 136*, 92-99. doi:10.1016/j.drugalcdep.2013.12.012

Connor, J. P., Haber, P. S., & Hall, W. D. (2016). Alcohol use disorders. *Lancet, 387*(10022), 988-998. doi:10.1016/S0140-6736(15)00122-1

DiBartolo, M. C., & Jarosinski, J. M. (2017). Alcohol Use Disorder in Older Adults: Challenges in Assessment and Treatment. *Issues in mental health nursing, 38*(1), 25-32. doi:10.1080/01612840.2016.1257076

Emiliussen, J., Nielsen, A. S., & Andersen, K. (2017). Identifying Risk Factors for Late-Onset (50+) Alcohol Use Disorder and Heavy Drinking: A Systematic Review. *Subst Use Misuse, 52*(12), 1575-1588. doi:10.1080/10826084.2017.1293102

Fingerhood, M. (2000). Substance Abuse in Older People. *Journal of the American Geriatrics Society, 48*(8), 985-995. doi:10.1111/j.1532-5415.2000.tb06900.x

Han, B. H., Moore, A. A., Sherman, S., Keyes, K. M., & Palamar, J. J. (2016). Demographic trends of binge alcohol use and alcohol use disorders among older adults in the United States, 2005–2014. *Drug and Alcohol Dependence, 170*, 198-207. doi:10.1016/j.drugalcdep.2016.11.003

Hasin, D. S., O'Brien, C. P., Auriacombe, M., Borges, G., Bucholz, K., Budney, A., . . . Grant, B. F. (2013). DSM-5 Criteria for Substance Use Disorders: Recommendations and Rationale. *The American journal of psychiatry, 170*(8), 834-851. doi:10.1176/appi.ajp.2013.12060782

Heron, J. E., Croudace, T. J., Barker, E. D., & Tilling, K. (2015). A comparison of approaches for assessing covariate effects in latent class analysis. *2015, 6*(4), 15. doi:10.14301/llcs.v6i4.322

Kuerbis, A., Moore, A. A., Sacco, P., & Zanjani, F. (2016). *Alcohol and Aging: Clinical and Public Health Perspectives*. Cham: Springer International Publishing.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS Procedure for Latent Class Analysis. *Struct Equ Modeling, 14*(4), 671-694.

Lemke, S., & Moos, R. H. (2003). Outcomes at 1 and 5 years for older patients with alcohol use disorders. *J Subst Abuse Treat, 24*(1), 43-50.

McCutcheon, A. (1987). *Latent Class Analysis*. In. Retrieved from http://methods.sagepub.com/book/latent-class-analysis doi:10.4135/9781412984713

Miller, W. R., & Boca, F. K. D. (1994). Measurement of drinking behavior using the Form 90 family of instruments. In (Vol. 12, pp. 112-118). United States.

Nielsen, B., Nielsen, A. S., Lolk, A., & Andersen, K. (2010). Elderly alcoholics in outpatient treatment. *Dan Med Bull, 57*(11), A4209.

Nylund-Gibson, K., & Choi, A. (2018). *Ten frequently asked questions about latent class analysis*.

Nylund-Gibson, K., & Masyn, K. E. (2016). Covariates and Mixture Modeling: Results of a Simulation Study Exploring the Impact of Misspecified Effects on Class Enumeration. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 782-797. doi:10.1080/10705511.2016.1221313

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535-569. doi:10.1080/10705510701575396

Oslin, D. W., Pettinati, H., & Volpicelli, J. R. (2002). Alcoholism Treatment Adherence: Older Age Predicts Better Adherence and Drinking Outcomes. *The American Journal of Geriatric Psychiatry, 10*(6), 740-747. doi:10.1097/00019442-200211000-00013

Rinker, D. V., & Neighbors, C. (2015). Latent Class Analysis of DSM-5 Alcohol Use Disorder Criteria Among Heavy-Drinking College Students. *J Subst Abuse Treat, 57*, 81-88. doi:10.1016/j.jsat.2015.05.006

Satre, D. D., Mertens, J. R., Arean, P. A., & Weisner, C. (2004). Five-year alcohol and drug treatment outcomes of older adults versus middle-aged and younger adults in a managed care program. *Addiction, 99*(10), 1286-1297. doi:10.1111/j.1360-0443.2004.00831.x

Skevington, S. M., Lotfy, M., & O'Connell, K. A. (2004). The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. *Qual Life Res, 13*(2), 299-310. doi:10.1023/b:Qure.0000018486.91360.00

Swift, W., Slade, T., Carragher, N., Coffey, C., Degenhardt, L., Hall, W., & Patton, G. (2016). Adolescent Predictors of a Typology of DSM-5 Alcohol Use Disorder Symptoms in Young Adults Derived by Latent Class Analysis Using Data From an Australian Cohort Study. *J Stud Alcohol Drugs, 77*(5), 757-765.

Vermunt, J. K. (2010). Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Political Analysis, 18*(4), 450-469. doi:10.1093/pan/mpq025

Wieben, E. S., Nielsen, B., Nielsen, A. S., & Andersen, K. (2018). Elderly alcoholics compared to middle-aged alcoholics in outpatient treatment - 6-month follow-up. *Nordic journal of psychiatry*, 1-6. doi:10.1080/08039488.2018.1522373

Anders Milhøj

Nyheder i SAS Analytics 15.1

Department of Economics, University of Copenhagen

Øster Farimagsgade 5, DK-1353 København K

Anders.Milhoj@econ.ku.dk

I efteråret 2018 blev Analytical Products i den opdaterede version 15.1 sendt på markedet. Denne opdatering indeholder, som så mange gange før, interessante opdateringer af de analytiske programpakker inden for statistik, økonometri, operationsanalyse etc. Disse opdateringer er løsrevet fra opdateringer af det samlede SAS-program, så det er stadig Base SAS, version 9.4, der anvendes.

SAS's nyere analytiske releases

Version 9.4 af Base SAS med Analytical Updates 14.1, som udkom sommeren 2015, blev midt i november 2018 opdateret for tredje gang, nu til version 15.1 af Analytical Products. I dette indlæg fokuseres på denne version 15.1, der som sædvanligt indeholder en del flere nyheder, mens de tidligere opdateringer er beskrevet i mange tidligere symposieindlæg.

Kilderne til disse nyhedsoversigter er er SAS-hjælpen, som kan tilgås af alle - også uden en SAS-installation - via http://support.sas.com/, idet manualerne for SAS-pakkerne STAT, ETS. OR, QC er offentligt tilgængelige for alle - Google er din ven!

Øget tilgængelighed til SAS

Atter en gang vil jeg reklamere for **SAS on Demand for Academics**, der er en serverløsning, hvor brugeren afvikler SAS-sessionen online på en fjernt beliggende server; helt som i gamle dage, hvor mainframes var eneste mulighed – det er dog desværre ikke længere mulighed for at anvende hulkort! En lærer kan oprette kurser og derved stille data til rådighede for de studerende, der tilmelder sig kurset. Men der er også mulighed oprette sig selv som "individual learner", så hvem som helst kan faktisk afvikle SAS-sessioner af den vej. Ikke alle hjørner af SAS er med, men hele STAT pakken og hele ETS pakken er heldigvis 100% med, så SAS on Demand dækker behovet for alle mine SAS-kurser.

For yderligere oplysninger om SAS on Demand, se

https://www.sas.com/da_dk/software/on-demand-for-academics.html

SAS Viya, Big Data, Python og andet nymodens gøgl

Listen over HP (High Performance) procedurer, som kopierer de gængse SAS procedurer, udvides stadigt. De regner hurtigt, og de udnytter maskinfigurationen fuldt ud. Fx kan de regne multithreadet, hvis maskinen indeholder flere processorer eller beregningerne kan fordeles over geografiske adskilte CPU-er. De stilles til rådighed for almindelige SAS brugere i simple PC installationer. Derved kan visse ekstra raffinementer udnyttes af alle som fx, at der i PROC HPREG, i modesætning til i PROC REG, stilles en CLASS statement til rådighed. Det kræver specielle licenser at udnytte faciliteterne til distribueret kørsel af SAS-programmer.

Der er også "pakker" der hjælper med til at overflødiggøre kodning, fx SAS Visual Data Mining and Machine Learning

SAS Viya is an extension to the SAS platform that supports high-performance analytical data preparation, variable transformations, exploratory analysis, analytical modeling, integrated model comparison, and scoring.

Et citat fra et nyhedsbrev:

Python has become a popular programming language for both data analytics and mathematical optimization. With SAS Viya and its Python interface, Python programmers can use the state-of-the-art optimization solvers that SAS provides. This paper demonstrates an approach for Python programmers to naturally model their optimization problems, solve them by using SAS® Optimization solver actions, and view and interact with the results. The common tools for using the optimization solvers in SAS for these purposes are the OPTMODEL and IML procedures, but programmers more familiar with Python might find this alternative approach easier to grasp.

STAT nyt

- NEW! **PROC BGLIMM** provides full Bayesian inference for generalized linear mixed models (GLMMs).
- NEW! **PROC CAUSALGRAPH** examines the structure of graphical causal models.
- NEW! **PROC RMSTREG** analyzes time-to-event data by using regression with respect to the restricted mean survival time (RMST).

- **PROC ICPHREG** fits the semiparametric proportional hazards model to interval-censored data.
- **PROC PHREG** performs Bayesian analysis of the proportional hazards spline model.
- **PROC QUANTREG** enables you to perform observation-wise conditional distribution analysis

ETS Nyt

SAS/ETS 15.1 includes a new interface engine and enhancements to several procedures:

- **SASEOECD interface engine**: This new engine enables you to retrieve time series data from the Organisation for Economic Co-operation and Development (OECD) website.
- **PROC AUTOREG**: The CLASS statement is production, and you can output standard errors of prediction intervals by specifying the STDERR= and STDERRM= options in the OUTPUT statement.
- **PROC COPULA**: You can use a SIMULATE statement with a BY statement, provided that you specify a FIT statement before the SIMULATE statement. You can also use multiple FIT and SIMULATE statements with a BY statement.
- **PROC COUNTREG**: This procedure includes two new options for variable selection. The NOSPLITEFFECTS option requests that effects that involve class variables not be split into individual effects that correspond to class levels, and the RETAINEFFECT option requests that the effects named within parentheses be retained during the variable selection process.
- **PROC SPATIALREG**: Taylor and Chebyshev approximation techniques for spatial error models and spatial Durbin error models have been added.
- **PROC SSM**: You can now request that parameter estimation be based on a new type of likelihood called marginal likelihood.
- **PROC TMODEL**: Most features are now production. Only the RANDOM statement and the QUADHESS=ANALYTIC option remain experimental. The LUSOLVER= option has been added to provide better control over the solution of matrix equations that occur in FIT and SOLVE tasks.
- **PROC UCM**: You can now add a transfer-function component to your model by using the new TF statement. You can also request that parameter estimation be based on a new type of likelihood called marginal likelihood.

- **PROC VARMAX**: Conditional forecasts and scenario analysis are supported for the vector autoregressive (VAR) model, Bayesian VAR model, vector error correction model (VECM), and Bayesian VECM, with or without exogenous variables. You can now plot dynamic conditional covariances for DCC GARCH models by specifying PLOTS=CONDCORR. The OUTSTAT statement now includes the value of the log-likelihood function that is calculated at the parameter estimates.

IML Nyt

- The CONTINUE statement stops the processing of the current iteration of a DO loop and resumes processing at the next iteration of the DO loop.
- The new digital filtering functions enable you to design digital filters and apply those filters to signals. Many of the digital filtering functions start with the "DF" prefix. The new function and subroutines are CCEPSTRUM, DFCONV, DFDESIGN, DFFILT, DFFREQZ, DFFREQZZPK, DFMEDFILT, DFORDER, DFSOSFILT, DFSOSFREQZ, DFSOSFREQZZPK, ICCEPSTRUM, and RCEPSTRUM.
- The experimental KPCATRAIN subroutine computes a kernel principal component (kPCA) analysis from training data. The experimental KCPASCORE function uses the kPCA model to score new data.
- The FEVAL function enables you to evaluate a function indirectly by specifying the name of the function and its arguments.
- The LEAVE statement exits the current DO loop and resumes processing at the statement that follows the DO loop.
- The MODULESTACK function returns the names of all modules in the module call stack.
- The SPECTROGRAM subroutine displays a spectrogram of a short-time Fourier transform of a time series signal.
- The TABLESORT subroutine sorts a table by one or more columns.
-

QC Nyt

- **PROC CAPABILITY** has the following enhancements:

o The BARFILL= option in the HISTOGRAM statement enables you to fill the histogram bars in different cells of a comparative histogram with different colors.

o The PCTLAXIS option in the QQPLOT statement has a new PCTLORDER= suboption, which specifies the tick mark values to be labeled on the percentile axis.

- In the INTERVALS statement, you can specify METHOD=7 to compute nonparametric tolerance intervals.

- **PROC RAREEVENTS** enables you to apply tests for special causes to rare events charts. These tests can improve the sensitivity of rare events charts by detecting patterns of measurements that might indicate unusual variation.

- The CCHART, NPCHART, PCHART, and UCHART statements in **PROC SHEWHART** support the new LANEY option, which computes control limits as recommended by Laney.

- The EWMACHART and MACHART statements in **PROC MACONTROL** support the new LABASYMPTOTIC option, which labels varying control limits with asymptotic limit values.

- **PROC SHEWHART**, **PROC ANOM**, **PROC CUSUM**, and **PROC MACONTROL** recognize new macro variables that you can use to control details of ODS Graphics output more conveniently.

OR Nyt

SAS/OR 15.1 includes performance improvements to its optimization procedures and solvers, including the following enhancements:

- The LP, MILP, network, QP, NLP, and CLP solvers and the decomposition algorithm (for LP and MILP) add performance and stability improvements.

- The LP and MILP solvers significantly reduce their memory consumption.

- The LP solver replaces its presolver.

- The MILP solver and the OPTMILP procedure add the ability to report multiple solutions.

- The network solver adds a path enumeration algorithm to find all paths between specified nodes.

- The connected components algorithm in the OPTNET procedure adds support for the thin internal graph format.

- The DECOMP algorithm adds the COMMUNITY value to the METHOD= option.

- The CLP solver adds the CUMULATIVE predicate, and the CLP procedure adds the CUMULATIVE statement.

- The NLP solver adds a new interior point solution algorithm on an experimental basis.

- The LSO (local search optimization) solver is added and is called by the OPTMODEL procedure.

PROC RAREEVENTS i QC pakken

I dette afsnit fortsættess eksmelet fra 2016 symposieindlægget på et datasæt, der bla indeholder eksakt tidspunkt, på sekundniveau, samt styrke for alle jordskælv i Oklahoma. Datasættet er downlaoded fra

htttp://earthquake.usgs.gov

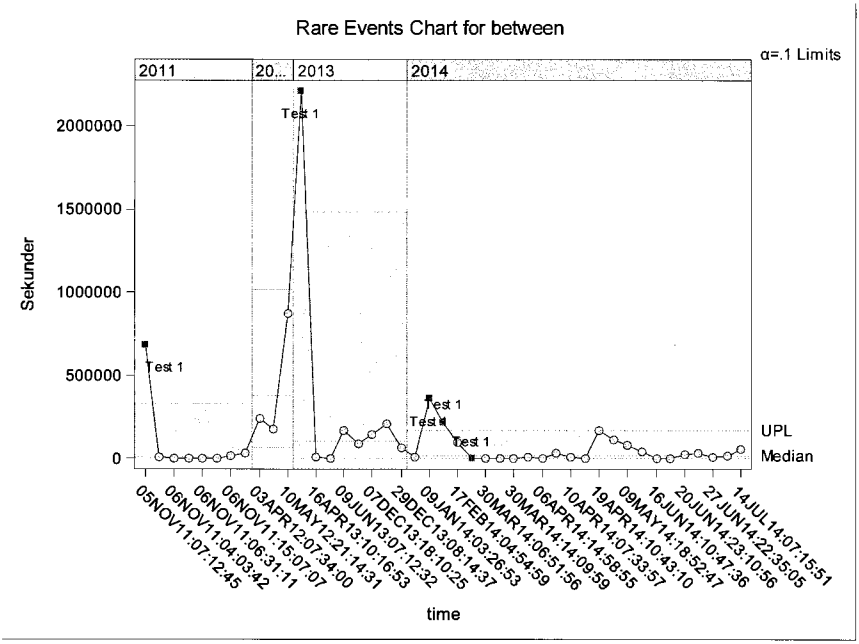
Nogle påstår at den udbredte udvinding af olie med den nye "fracking" metode leder til flere jordskælv. Dette kan let illustreres med PROC RAREEVENTS.

I programmet analyseres data fra perioden 1. januar 2010 til ca 1/10 2015, idet der betragtes jordskælv af styrke over 3.6. For at illustrere de nyere faciliteter er tidsperioden inddelt faser svarende til et kalenderår. Desuden er optionerne for test1 - test4 for usædvanlige begivenheder tilføjet. Desuden angives grænserne ved en Weibull fordeling.

```
PROC RAREEVENTS DATA=wrk8.Oklahoma;
   COMPARE BETWEEN/NBINS=45;
   CHART BETWEEN * time/ALPHALPL=0.05 ALPHAUPL=0.05
   dist=Weibull test1 test2 test3 test4
   readphases=all phaselegend limitphases=all;
   LABEL BETWEEN = 'Sekunder';
   ID time;
   where year(datepart(time))>2010 and mag>3.6;
RUN;
```

Output fra proceduren består en række plots, hvoraf et enkelt er vist nedenfor. Plottet viser data fra 2011, 2012, 2013 og det første halvår af 2014, idet andet halvår og de første tre kvartaler af 2015 tilsammen fylder to tilsvarende plots.

Anden aksen på figuren er angivet i sekunder, så tallene er store. Det ses, at der har været lange tidsintervaller mellem jordskælvene i de første år – op mod halve år. Efter omkring årsskiftet 2013/2014 optræder jordskælvenen langt hyppigere. Husk at hvert punkt på kurven repræsenterer et enkelt jordkælv og ikke kalendertiden, så når der lange intervaller for 2014 og 2015 er fordi, der er mange jordskælv. Punkter markeret med "test1" er tidsintervaller mellem jordskæv, der falder udenfor 95% konfidensintervallet, der er bestemt ud fra Weibull fordelingen. De øvrige muligheder for testoptionerne, fx mange observationer i træk over eller under medianen i Weibull fordelingen, optræder ikke i denne graf



PROC RMSTREG i STAT pakken

Teori

The procedure analyzes time-to-event data by using regression with respect to the restricted mean survival time (RMST).

Let T be a nonnegative random variable that represents the failure time of an individual from a homogeneous population. The survival function (also known as the survivor function) of T is defined as

$S(t)=\Pr(T\geq t)$

Assume that τ is a prespecified time point of interest. Let R be the minimal of T and τ, then the restricted mean survival time (RMST) is defined as the expected value of R:

$\text{RMST}(\tau)=E(R)=E[\min(T,\tau)]$

It can be evaluated as the area under the survival function, which is estimated numerically, corrected for the effect of censoring at time τ.

When the largest observed time is censored, the mean survival time is underestimated because the Kaplan-Meier estimate does not drop to zero (Klein and Moeschberger 2003). The RMST, on the other hand, can still be estimated unbiasedly as long as τ is less than the largest observed time to event.

Under the natural logarithm link, the model is

$$\log[\mathrm{RMST}_i(\tau)] = \mathbf{x}_i^T\boldsymbol{\beta}$$

Pseudovalue regression is a generic method of fitting generalized linear models to time-to-event data (Andersen, Klein, and Rosthøj 2003). Andersen, Hansen, and Klein (2004) proposed using pseudovalue regression to analyze the RMST models. Because the nonparametric estimator is unbiased, it can be used in place of in the estimation process. This method is default in PROC RMSTREG.

Et eksempel

I eksemplet er denne procedure anvendt på datasættet for polit studerende, der blev anvendt i et symposieindlæg i 2018. Data omfatter alle, der begyndte bachelorstudiet sommeren 2015. Ud af disse omkring 300 studerende er 49 faldet fra inden for den tid datasættet strækker over, dvs før sommereksamen 2016. I analysen anvendes forklarende variable, der er til rådighed ved studiestart, fx alder, geografiske oplysninger og oplysninger om den foregående ungdomsuddannelse. Desværre mangler oplysningerne dog for visse af de studerende i datamaterialet.

En analyse med et udvalg af forklarende variable afvikles ved følgende program.

```
proc rmstreg data=analyse tau=275;
   class rustur;
   model dage*fiasko(0) = rustur start_alder ADGex_kvotient
p_postnr/ link=linear method=pv;
run;
```

Output fra dette procedurekald viser, at den adgangsgivende kvotient er mest signifikant, dvs at tiden før frafald kan forventes at stige jo højere adgangskvotienten er. Alderen ved studieår har også en vis betydning, så varigheden før frafald er lavest for ældre studiestartere. Det fremgår tydelige hvis årstallet for den adgangsgivende eksamen benyttes om en class variabel, idet studerende med helt frisk adgangsgivende eksamen fra 2015 flader signikant hurtigere fra end andre.

Analysis of Parameter Estimates								
Parameter		DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	225.4136	16.4748	193.1236	257.7037	187.21	<.0001
rustur	ikke rustur	1	2.1607	5.9074	-9.4176	13.7389	0.13	0.7145
rustur	rustur	0	0.0000					
start_alder		1	0.0026	0.0015	-0.0003	0.0055	3.20	0.0738
ADGEX_KVOTIENT		1	2.7015	0.9961	0.7492	4.6537	7.36	0.0067
p_postnr		1	-0.0025	0.0023	-0.0070	0.0021	1.13	0.2871

I analysen indgår postnummer med sin numeriske værdi, da postnummeret i en vis forstand viser afstanden fra "barndomshjemmet" til Frue Plads, men denne forklarende variable har ingen effekt. I stedet kan skolen navn (fx gymnasiets navn) bruges som class variable. Samlet set giver det en signifikant effekt med p = 2.1%, men da der utroligt mange gymnasier etc, der har leveret studerende til økonomistudiet, kan listen over estimater ikke bruges til noget.

I stedet bruges første ciffer i postnummeret, idet det angiver den adresse den studerende havde, da han/hun søgte optagelse. I tabellen benyttes et 5-cifret postnummer med begyndelse 24 som grundlag. Dette postnummer er i Flensborg, men det ses af tabellen, at studerende fra Fyn (5000) og Jylland (Omkring Århus, 8000) giver samme estimationsresultat. Studerende fra Region Sjælland (gruppe 4000) er ikke signifikant længere om at falde fra. Studerende fra postnummer i gruppe 2000 (København uden for Indre By samt forstæderne) og 3000 (Nordsjælland, men ikke deciderede forstæder) falder signifikant hurtigere fra end jyder og fynboer.

Analysis of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	275.0000	0.0000	275.0000	275.0000	1.3E14	<.0001
postnummer_gruppe 1	1	-4.7857	3.3882	-11.4266	1.8551	2.00	0.1578
postnummer_gruppe 2	1	-9.4233	2.4670	-14.2586	-4.5879	14.59	0.0001
postnummer_gruppe 3	1	-22.8148	10.8123	-44.0065	-1.6231	4.45	0.0349
postnummer_gruppe 4	1	-17.4375	11.6810	-40.3318	5.4568	2.23	0.1355
postnummer_gruppe 5	1	-0.0000	0.0000	-0.0000	0.0000	0.00	1.0000
postnummer_gruppe 8	1	-0.0000	0.0000	-0.0000	0.0000	0.00	1.0000
postnummer_gruppe 24	0	0.0000					

Referencer

Andersen, P. K., Hansen, M. G., and Klein, J. P. (2004). "Regression Analysis of Restricted Mean Survival Time Based on Pseudo-observations." *Lifetime Data Analysis* 10:335–350.

Andersen, P. K., Klein, J. P., and Rosthøj, S. (2003). "Generalised Linear Models for Correlated Pseudo-observations, with Applications to Multi-state Models." *Biometrika* 90:15–27.

Diverse hjemmesider på support.sas.com. Passager på engelsk er direkte citater.

Januar

2019

Undersøgelse af karakterer i matematik og mikroøkonomi på HA/BA-uddannelserne ved Syddansk Universitet

Nils Karl Sørensen nks@sam.sdu.dk
Institut for Virksomhedsledelse og Økonomi
Syddansk Universitet

Resumé

Kurset i mikroøkonomi på HA/BA-uddannelserne er typisk placeret på 1. semester og kan indholdsmæssigt gøre brug af matematiske metoder til løsning af kursets temaer, hvilket mange studenter finder vanskeligt. Ofte kan dette resultere i dårlige eksamensresultater og deraf følgende frafald.

Disse temaer behandles med udgangspunkt i en population af 2.425 eksamensforsøg i mikroøkonomi kaldet virksomhedens økonomi I og II ved Syddansk Universitet. Materialet undersøges i forhold til den gymnasiale adgangsgivende karakter, matematikkarakteren, den gymnasiale uddannelsesbaggrund samt køn.

Generelt findes det, at studenterne klarer sig dårligere til eksamen i VØ I og II end til den gymnasiale eksamen i matematik. Dette afhænger af såvel køn som uddannelsesbaggrund. Det findes også, at studenter med en dårlig karakter i matematik har en større motivation for at arbejde, så de får et bedre resultat i VØ I og II.

1. Indledning

Undervisningen i mikroøkonomi giver ofte anledning til bekymring i studienævn for erhvervsøkonomi ved Syddansk Universitet (SDU). Kurset, der på SDU benævnes virksomhedens økonomi I og II (VØ I og VØ II) er placeret på HA-studiets 1. og 2. semester. Kurset betjener sig indenfor nogle af emnekredsene af matematiske tilgange, der ofte kan volde studenterne besvær på trods af, at disse på 1. semester følger et kursus i matematik, hvor der ofte eksemplificeres med problemstillinger i undervisningen i VØ I og II. Til eksamen i VØ I og VØ II er der ofte mange studenter, der dumper. Det er en dårlig begyndelse på studiet og kan lede til, at en studerende

måske opgiver studiet på et fejlagtigt grundlag. På denne baggrund kan følgende problemstillinger forekomme relevante at undersøge:

- Er der en sammenhæng mellem karakteren i VØ I og VØ II og den adgangsgivende gymnasiale karakter i matematik?
- Er der en sammenhæng mellem karakteren i VØ I og II og den gymnasiale specialisering (STX, HHX med mere)?
- Hvordan klarer studenterne med eksempelvis en dårlig gymnasial karakter i matematik sig specifikt i VØ I og VØ II?
- Har det adgangsgivende karaktergennemsnit en afvigende indflydelse på bestået af VØ I og VØ II i forhold til den gymnasiale karakter i matematik?
- Hvordan vil eventuelle mindstekrav i karakteren i matematik fra gymnasiet øve indflydelse på sandsynligheden for at bestå VØ I og VØ II?

Artiklen er organiseret som følger. Først præsenteres undersøgelsens statistiske materiale. Dernæst ses der på sammenhængen til den adgangsgivende karakter i matematik. I afsnit 4 relateres matematikkarakteren til karaktererne i VØ I og VØ II, mens afsnit 5 indeholder en forløbsanalyse af sammenhængen. I afsnit 6 ses der på relationen til den adgangsgivende karakter, mens afsnit 7 er sammenfatning og konklusion.

2. Det statistiske materiale

Materialet indeholder oplysninger for population på 2.425 eksamensforsøg (studenter) omkring resultater i de mikroøkonomiske fag Virksomhedens Økonomi I (VØ I) og Virksomhedens Økonomi II (VØ II) samt resultaterne i matematik opnået på den adgangsgivende uddannelse.

Studenterne har haft mindst én eksamensrelateret aktivitet i de to kurser VØ I og VØ II. Materialet dækker optag fra 2014 til 2016, og er trukket fra Syddansk Universitets database primo april 2017.

Materialet er fordelt på ordinær og reeksamen (første og andet forsøg) samt på studentens adgangsgivende gymnasiale uddannelse. Oplysningerne om matematik fra den adgangsgivende eksamen er fordelt på henholdsvis A- og B-niveau.

Af de 2.425 studenter er det muligt at danne et konsistent materiale for 2.106 studenter. Alle disse er optaget via KOT. Restgruppen er på 2 genoptagne, 5 overflyttede fra anden institution samt 312 studenter optaget som "indskrevet til afrundet uddannelse (som led i anden uddannelse eksempelvis indskrevet til en

supplerende uddannelse)". For disse optagne findes der ikke oplysninger for den adgangsgivende eksamen i matematik.

Med hensyn til matematik, så foreligger oplysninger for matematik på såvel A- som på B-niveau. Der knytter sig flere problemer til materialet. Der foreligger såvel årskarakter, eksamenskarakter, mundtlig og skriftlig karakter. Det er imidlertid ikke klart, om den mundtlige og/eller skriftlige karakter er opnået til eksamen eller via årskarakteren. På STX-uddannelsen (almene gymnasium) gives der såvel års- som eksamenskarakterer, mens der på HF-uddannelsen alene gives karakterer for eksamenspræstationen. Der kan også være problemer forbundet med at konvertere karakterer for internationale studenter. Endelig virker praksis omkring indberetning ikke konsistent mellem uddannelsesinstitutioner.

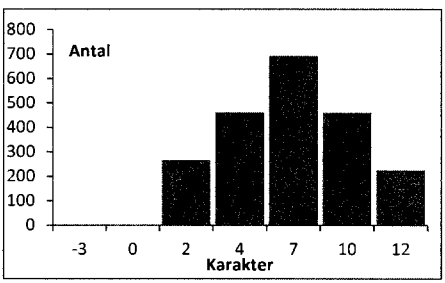
For at løse denne problemstilling, er der konstrueret en variabel for den enkelte students samlede formåen i matematik. Der kan være to tilgange hertil. Dels kan man tage et gennemsnit af alle opnåede resultater dels kan man vælge en karakter, som anses som repræsentativ for studentens formåen i matematik. Der er to ulemper ved at vælge gennemsnittet. For det første bliver det meget vanskeligt at opnå karakteren 12. For det andet kan man risikere, at eksempelvis dobbelttastninger medgår i karakteren. Anvendes gennemsnittet bliver variationen i materialet også mindre. Den anden metode er, at vælge en karakter. I denne undersøgelse er valgt at anvende den højeste karakter. Det kan give en positiv bias, men omvendt får man en variabel med større variation. Som det vil fremgå af de følgende afsnit, vurderes denne bias til at være beskeden.

3. Den adgangsgivende karakter i matematik

Tabel 1 gengiver resultatet af studenternes adgangsgivende resultat i matematik. Karaktererne fordeler sig symmetrisk omkring karakteren 7 med lidt flere 4- og 7-taller end idealfordelingen som givet af karakterskalaen med procentfordeling som 10, 25, 30, 25 og 10.

Tabel 1: Adgangsgivende karakter i matematik

Karakter	Frekvens	% fordeling
-3	0	0.0
0	2	0.1
2	266	12.6
4	461	21.9
7	692	32.9
10	460	21.8
12	225	10.7
Sum	2,106	100.0



Tabel 2 og 3 giver yderligere informationer om den adgangsgivende karakter i matematik, idet materialer er fordelt på en række baggrundsparametre

Tabel 2: Karakterer i matematik på niveau og køn

	Alle	MAT-A	MAT-B	Kvinder	Mænd
Middelværdi	6.9	7.2	6.8	7.1	6.8
Median	7	7	7	7	7
Typetal	7	7	7	7	7
Standardafvigelsen	3.1	3.0	3.2	3.2	3.1
Observationer	2,106	576	1,530	807	1,299

Studenterne med matematik A har bedre resultater end matematik B. Over 70 procent af studenterne komme med matematik på B niveau. Det kan have betydning for den pædagogiske tilgang samt omkring valg af lærebog. Ikke kun i VØ I og VØ II, men også i matematikkurset på HA/BA-uddannelsen. Der er 38 procent kvinder, og de klarer sig 0.3 point bedre end mændene med hensyn til karakteren. Denne forskel er netop signifikant ved et 5 procents niveau.

Tabel 3: Karakter i matematik på type af adgangsgivende uddannelse

	UOP	STX	INT	HHX	HTX	HF	ANDU
Middelværdi	3.0	7.1	3.3	7.5	6.0	7.1	7.8
Median	2	7	2	7	7	7	8.5
Typetal	2	7	2	7	7	7	IM
Standardafvigelsen	2.3	2.8	2.9	3.0	2.6	3.1	4.3
Observationer	66	734	113	874	66	249	4

Note: UOP = Uoplyst, INT = International, ANDU = Anden uddannelse, IM = Ikke mulig at beregne

Studenter, der kommer med en HHX-uddannelse har den bedste adgangsgivende karakter i matematik, når der ses bort fra de 4 studenter med anden uddannelse. Andelen af HHX'ere er på 41.5 procent og er den højeste. HHX er nok den gymnasiale uddannelse, der bedst matcher HA/BA-uddannelsen. Niveauerne for STX og HF er identiske, mens karakteren på HTX er noget lavere. Andelen af STX'ere er den næststørste og andrager næsten 35 procent. For internationale studenter samt gruppen "ikke oplyst" er karakteren i matematik væsentligt lavere på henholdsvis 3.3 og 3.0. Dette kan måske henføres til konvertering af karakterer fra udlandet til Danmark. De ovennævnte tendenser afspejler sig også i medianen og typetallet. Det bemærkes, at variationen er konstant for alle kategorier.

4. Resultater af eksamen i VØ I og VØ II

Tabellerne 4, 5 og 6 sammenfatter nogle hovedresultater af eksamensresultaterne for eksamen i VØ I og VØ II. Sammenligning med tabel 3 og 4 viser, at resultaterne i VØ I og VØ II er markant lavere end resultatet i matematik fra gymnasiet.

Tabel 4: Resultater og forsøg af eksaminer i VØ I og VØ II

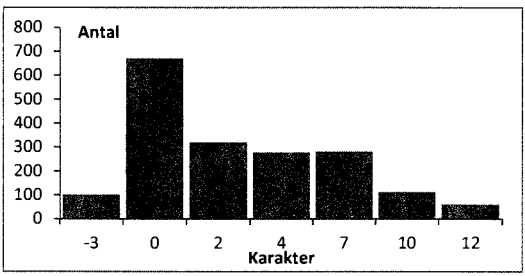
	VØ I 1	VØ I 2	VØ II 1	VØ II 2
Middelværdi	2.9	1.7	2.5	1.7
Median	2	0	2	0
Typetal	0	0	0	0
Standardafvigelsen	3.7	2.5	4.1	2.9
Observationer	1,823	729	988	441

Note: 1 = forsøg 1, 2 = forsøg 2

Endvidere er resultatet i VØ II markant dårligere end i VØ I. Dette kan skyldes, at der er progression i fagets indhold. For det andet, er resultatet i andet forsøg (reeksamen) signifikant lavere end resultatet i første forsøg. Endelig er variationen i første forsøg større end i andet forsøg. Dette forhold kan henføres til en større dumpeprocent i første forsøg, og det forhold, at færre får topkarakteren i andet forsøg.

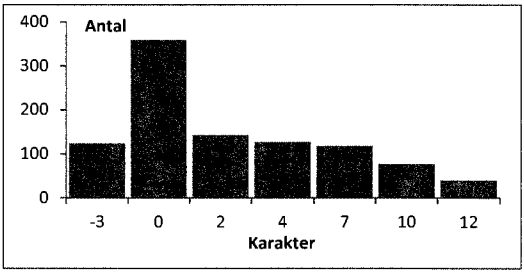
Tabel 5: Resultater VØ I første forsøg

Karakter	Frekvens	% fordeling
-3	101	5.5
0	671	36.8
2	320	17.6
4	278	15.2
7	281	15.4
10	112	6.1
12	60	3.3
Sum	1,823	100.0



Tabel 6: Resultater VØ II første forsøg

Karakter	Frekvens	% fordeling
-3	124	12.6
0	359	36.3
2	143	14.5
4	127	12.9
7	118	11.9
10	77	7.8
12	40	4.0
Sum	988	100.0



Histogrammerne i tabel 5 og 6 viser at karakteren 0 generelt er den hyppigste karakter i begge fagene. Fordelingerne af resultaterne er generelt identiske og for de bestående faldende med stigende karakter. I VØ II er der flere, der får -3 end i VØ I. Dette trækker gennemsnittet ned. Andelen af studenter med 10 eller 12 i VØ II er således større, mens andelene med 2, 4 og 7 er mindre.

Tabellerne 7 og 8 viser resultater i VØ I og VØ II fordelt efter den adgangsgivende gymnasiale uddannelse. Bemærk, at der i nogle af tilfældene, er få observationer. Det genfindes generelt, at resultaterne til reeksamen er signifikant svagere end til den ordinære eksamen.

Tabel 7: Resultater og forsøg af eksaminer VØ I fordelt efter gymnasial uddannelse

	UOP	STX	INT	HHX	HTX	HF
	F1/F2	*F1/F2*	*F1/F2*	*F1/F2*	*F1/F2*	*F1/F2*
Middelværdi	5.2/1.5	2.7/1.8	3.7/1.8	2.9/1.8	4.7/1.4	1.5/1.1
Median	4/0	2/0	4/2	2/2	4/0	0/0
Typetal	7/0	0/0	0/0	0/0	7/0	0/0
Standardafvigelsen	4.1/2.1	3.6/2.6	4.0/2.6	3.7/2.5	4.2/2.8	3.0/2.4
Observationer	60/13	657/280	99/40	770/288	57/14	176/90

Note: F1 = forsøg I, F2 = forsøg II

Studenter med uoplyst bagrund, international baggrund samt HTX klarer sig bedre end det samlede gennemsnit for første eksamensforsøg i VØ I på 2.9. Perspektiveres dette i forhold til adgangskravet i matematik i tabel 3, så er dette overraskende, da disse

studenter havde de svage adgangskarakterer i matematik. Af de danske studenter fandtes også det omvendte for HTX, hvor gennemsnittet er på 4.7.

Af de øvrige gymnasiale uddannelser har alle i første forsøg gennemsnit under 3. Bedst er HHX fulgt af STX, mens HF står svagest.

Sammenhængen mellem den gymnasiale karakter og resultatet i VØ I er således svag, når det gælder den gymnasiale uddannelsesbaggrund. Der kan være flere årsager til resultatet. Gruppen af uoplyst uddannelse kan mere heterogen. Standardafvigelsen for denne gruppe er også stor. For de internationale studenter kan to forhold være gældende. For det første kan det tænkes, at det er vanskeligt, at vurdere de foreliggende kundskaber i matematik, hvor disse kan være undervurderet. For det andet kan det tænkes, at disse studenter har en mere seriøs tilgang til faget, måske betinget i ringe økonomiske forhold, og dermed lav indkomst i de lande, hvorfra studenterne kommer.

Ser man på forsøg II, så er resultaterne ikke overraskende svagere. Til dette eksamensforsøg er det de internationale studenter, HHX og STX, der klarer sig bedst alle med middelværdier på 1.8. Der fremkommer således ikke et entydigt mønster i forhold til den ordinære eksamen.

Tabel 8: Resultater og forsøg af eksaminer VØ II fordelt efter gymnasial uddannelse

	UOP	STX	INT	HHX	HTX	HF
	F1/F2	*F1/F2*	*F1/F2*	*F1/F2*	*F1/F2*	*F1/F2*
Middelværdi	5.0/2.1	1.9/1.9	3.7/1.3	2.8/1.7	5.4/1.9	0.8/0.7
Median	4/2	0/0	4/0	2/2	7/2	0/0
Typetal	0/0	0/0	0/0	0/0	7/0	0/0
Standardafvigelsen	4.5/2.5	3.9/3.0	4.2/2.9	4.1/3.0	4.6/2.9	3.1/2.7
Observationer	41/9	347/173	70/15	416/183	29/14	80/46

Note: F1 = forsøg I, F2 = forsøg II

Resultaterne for VØ II findes, som nævnt i tabel 8. Her genfindes mange af tendenserne fra VØ I om end på et lavere karakterniveau. De samme tre grupper, som tidligere, er de bedste. Det bemærkes, at studenter med en HF baggrund er meget svage i VØ II. Da det generelle karakterniveau i VØ II er lavere har studenterne selvfølgelig et mere identisk resultat til eksamensforsøg II.

5. Forløbsanalyse

I det følgende undersøges, hvordan studenter med en bestemt karakter i matematik forventes at klare sig i henholdsvis VØ I og VØ II. For nærværende er resultaterne for såvel ordinær som reeksamen i de to respektive kurser slået sammen. Efter korrektion for manglende resultater og frafald, kan der konstrueres et panel på 1,084 eksamensresultater i forløb.

Tabel 9: Forløbsanalyse af gennemsnit VØ I og VØ II

Mat	VØ I	VØ II	n	VØ I - Mat	VØ II - Mat
2	3.9	3.4	167	1.9	1.4
4	2.3	1.5	207	-1.7	-2.5
7	3.3	2.6	358	-3.7	-4.4
10	4.4	4.2	236	-5.6	-5.8
12	6.4	5.8	116	-5.6	-6.2
Total	3.8	3.2	1,084		

Note: Der er set bort fra 2 studenter, der har fået 0 i matematik

Resultatet er gengivet i tabel 9. Som udgangspunkt må det forventes, at studenter med karakteren 2 vanskeligt kan klare sig ringere end den gymnasiale karakter. Omvendt gør sig gældende for studenter, der har en gymnasial karakter på 12 i matematik. Disse kan klare sig uændret eller ringere. Da fordelingen af karakterer i VØ I og VØ II er åben nedadtil, må det forventes, at faldet for studenter med 12 i matematik, er større end stigningen i resultat for studenter med et gymnasialt 2-tal i matematik.

Som tidligere er der et fald i karaktern i matematik til VØ I og et mindre fald mellem VØ I og VØ II. Korrelationen mellem de to sidstnævnte karakterer er således den højeste og andrager 0.6, mens korrelationen til matematik for begge fag ligger mellem 0.23 og 0.25. Dette er som forventet.

Studenterne med en adgangsgivende eksamen i matematik på 2 klarer sig rigtig godt i analysen og bedre end studenter, der har en adgangsgivende karakter i matematik på 4 eller 7. Alle studenter med karakteren 4 eller højere klarer sig dårligere i fagene i forhold til karakteren i matematik, som det fremgår af de to kolonner til venstre. Her er karakteren i matematik fratrukket henholdsvis karakteren i VØ I og VØ II. Er differencen negativ klarer studenten sig dårligere end til eksamen i matematik. En forklaring på resultatet for karakteren 2 kunne være at studenterne, der er svage i matematik, godt er klare over dette. Derfor kunne man forestille sig, at disse havde en mere seriøs tilgang kurserne og dermed bedre resultater. Sammenlignes der med histogrammet for matematikkarakteren er der også præcis 100 færre studenter med karakteren 2, som kommer igennem hele forløbet svarende til 27.5 %. Dette er

imidlertid ikke enestående, da det tilsvarende tal for studenterne med karakteren 12 er på 48 procent. Generelt klarer studenterne sig bedre i VØ I end i VØ II, som tidligere vist. Har man en tocifret karakter i matematik klarer man sig generelt bedre. Har man 10 eller 12 i matematik opleves dog et betydeligt fald i den opnåede karakter i såvel VØ I som i VØ II. For studenter med karakteren 10 i matematik er reduktionen på henholdsvis 5.6 points i VØ I og 5.8 points i VØ II. For studenter med karakteren 12 i matematik er de tilsvarende værdier 5.6 og 6.2 points.

Tabel 10 uddyber yderligere resultaterne i VØ I og II anskuet i forhold til karakteren i matematik. Tabellen og grafikken skal læses på følgende måde: Alle resultater i VØ I og II er sorteret i forhold til karakteren i matematik. Den grønne linje i tabellen indikerer, hvor stor en andel af studenterne, der holdt deres karakter i matematik til eksamen i VØ I og/eller VØ II. Dette ses også af den grønne søjle i grafikken. Søjlerne til højre viser således, hvilke resultater eller andele, der er bedre i VØ I og/eller VØ II, mens det omvendte gør sig gældende for søjlerne til venstre.

Indledningsvis bemærkes det, at andelene med karakteren -3 er lav. Dette kan skyldes, at den ordinære og reeksamen er slået sammen. Et -3 indikerer således, at man har opnået denne karakter til begge eksaminer. Sandsynligheden for denne hændelse vurderes som lav.

Generelt er andelene af studenterne, der fastholder sit niveau i matematik i forhold til resultatet i VØ I og VØ II, faldende med stigende karakter. Karakteren 10 er dog en undtagelse. Her fastholder kun 10.4 procent af studenterne deres karakter, mens de tilsvarende andele for karaktererne 7 og 12 er på henholdsvis 15.2 og 14.2 procent.

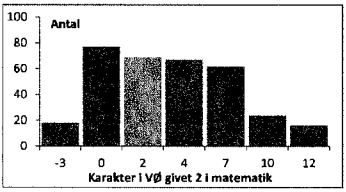
Fordelingerne i histogrammerne til højre i tabellen er forskellige. For matematikkarakteren 2 er der tale om en mere uniform fordeling, mens fordelingen blandt de beståede med matematikkarakteren 12 er mere symmetrisk omkring karakteren 7. Især histogrammet for karakteren 10 er ganske symmetrisk, mens fordelingerne for karakteren 4 og 7 er ganske højreskæve.

Af eksaminerne med matematikkarakteren 2 får næsten 5 procent 12, mens andelen er på godt 14 procent for studenter med matematikkarakteren 12. Det betyder, at hvis man har fået 12 i matematik på de gymnasiale uddannelser, så er der næsten 86 procents chance for et ringere resultat i VØ I eller VØ II. Har man matematikkarakteren 2, så er der omvendt over 50 procents chance for at man forbedre sit resultat til eksamen i VØ I eller VØ II. For karakteren 4 i matematik, er der næsten 40 procent, der får karakteren 0 VØ I eller VØ II. Dette kunne tyde på, at disse studenter ikke er så opmærksomme på udfordringerne i VØ I eller VØ II, som studenterne, der har en matematikkarakter på 2 med fra den gymnasiale uddannelse.

Tabel 10: Karakterer i VØ I og II givet bestemt karakter i matematik

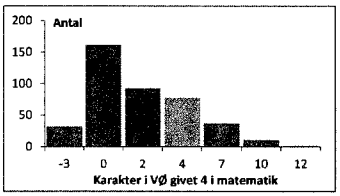
Karakter 2 i matematik

Karakter	Frekvens	% fordeling
-3	18	5.4
0	77	23.1
2	69	20.7
4	67	20.1
7	62	18.6
10	24	7.2
12	16	4.8
Sum	334	100.0



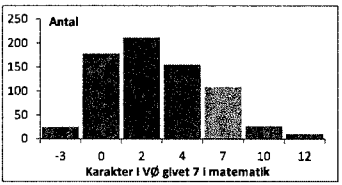
Karakter 4 i matematik

Karakter	Frekvens	% fordeling
-3	32	7.7
0	162	39.0
2	93	22.4
4	78	18.8
7	37	8.9
10	11	2.7
12	2	0.5
Sum	415	100.0



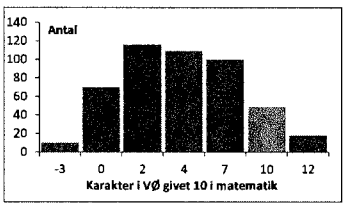
Karakter 7 i matematik

Karakter	Frekvens	% fordeling
-3	25	3.5
0	178	24.9
2	211	29.5
4	155	21.6
7	109	15.2
10	27	3.8
12	11	1.5
Sum	716	100.0



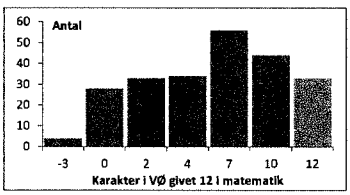
Karakter 10 i matematik

Karakter	Frekvens	% fordeling
-3	10	2.1
0	70	14.8
2	116	24.6
4	109	23.1
7	100	21.2
10	49	10.4
12	18	3.8
Sum	472	100.0



Karakter 12 i matematik

Karakter	Frekvens	% fordeling
-3	4	1.7
0	28	12.1
2	33	14.2
4	34	14.7
7	56	24.1
10	44	19.0
12	33	14.2
Sum	232	100.0



6. Undersøgelse af sammenhænge til adgangsgennemsnit

Materialet indeholder også informationer om gennemsnittet for den adgangsgivende gymnasiale eksamen. Indledningsvis betragtes tabel 11, der i opbygning ligner tabel 2 og 3 omhandlende karakteren i matematik.

Sammenholdes resultaterne med tabellerne omhandlende karakteren i matematik, så observeres der en række forskelle. Samlet samt fordelt på køn er det adgangsgivende karaktergennemsnit lavere end karakteren i matematik, men som tidligere findes det, at karaktererne for kvinder er højere end for mænd. At den adgangsgivende karakter er lavere, kan henføres til det forhold, at der er tale om en gennemsnitskarakter, der kan dække over en betydelig variation.

Tilsvarende gør sig gældende med hensyn til den adgangsgivende karakter fordelt på uddannelse med undtagelse af gruppen uoplyst samt for internationale studerende. Her er gældende, at matematikkarakteren er lav jævnfør tabel 3, mens det adgangsgivende karaktergennemsnit er højt. Disse grupper er således generelt gode, men svage når det gælder matematiske færdigheder. Dette kan forekomme lidt overraskende, da gruppen uoplyst klarer sig godt i såvel VØ I som i VØ II jævnfør tabellerne 7 og 8. De internationale studenter er med hensyn til VØ I og II placeret i midten.

For de gymnasiale uddannelser gælder, at HHX klarer sig bedst fulgt af STX, HTX og HF. Dette er som fundet tidligere.

Tabel 11: Adgangsgivende gennemsnit på køn og type af uddannelse

	Alle	Kvinder	Mænd	UOP	STX	INT	HHX	HTX	HF
Middelværdi	6.3	6.5	6.2	6.9	6.0	6.8	6.5	5.9	6.2
Median	6.2	6.4	6.1	7.2	5.9	6.7	6.4	6.0	6.1
Typetal	6.4	6.4	6.3	6.9	6.4	6.0	6.4	6.3	5.6
Standardafvigelsen	1.9	1.9	1.8	2.3	1.7	2.6	1.8	1.7	2.0
Observationer	2,058	788	1,270	51	734	84	940	66	249

Note: UOP = Uoplyst, INT = International

Da den adgangsgivende eksamen er et gennemsnit af mange enkeltpræstationer, kan denne opfattes som en kontinuert variabel. Tilsvarende kan der beregnes et udtryk for gennemsnittet af karakteren i VØ I g VØ II for de 1,047 individer, der har et resultat i begge disse kurser. For disse kan der opstilles en regressionsanalyse, hvor gennemsnittet i VØ I og VØ II kaldet VØ-SNIT kan forklares ved den adgangsgivende karakter samt dummyvariabler for køn samt typen af uddannelse. For nærværende er

de studerende, der har uoplyst adgangsgivende uddannelse anvendt som reference og inkluderet i konstantleddet.

Resultaterne af regressionsmodellen er vist i tabel 12. Modellens forklaringsgrad vurderes som tilfredsstillende set i lyset af karaktern af det statistiske materiale, og samtlige de forklarende variable er signifikante med undtagelse af koefficienten for dummyvariablen til HTX. Dette er en ret begrænset gruppe af studenter med 66 observationer. Da konstantleddet ikke er signifikant kan effekterne af de signifikante dummyvariabler vurderes direkte uden at studenter med uoplyst uddannelse eller HTX øver indflydelse på resultatet.

Tabel 12: Multipel regression af den gennemsnitlige karakter i VØ I og II

Regressions statistik	
Numerisk korrelation	0.55
R kvadrat	0.30
Justeret R kvadrat	0.30
Standardfejl	2.70
Observationer	1047

ANAVA

	fg	*SS*	*MS*	*F-test*	*P-værdi*
Regression	6	3309.97	472.85	64.86	0.00
Residual	1040	7574.37	7.29		
Total	1046	10884.34			

	Koefficient	*Std ofv.*	*t Stat*	*P-værdi*	*Lav 95 %*	*Høj 95 %*
Skæring	-0.12	0.59	-0.20	0.84	-1.27	1.03
ADG0VO	0.84	0.04	18.84	0.00	0.75	0.93
KØN	0.94	0.17	5.41	0.00	0.60	1.28
HF	-3.61	0.56	-6.42	0.00	-4.71	-2.51
HHX	-2.47	0.50	-4.91	0.00	-3.46	-1.48
HTX	-0.44	0.65	-0.68	0.50	-1.72	0.84
INT	-2.08	0.61	-3.39	0.00	-3.28	-0.88
STX	-2.25	0.51	-4.43	0.00	-3.24	-1.25

Det fremgår, at et godt adgangsgivende karaktergennemsnit har en positiv indflydelse på karakteren i VØ I og II. Man opnår typisk 84 procent af sit gennemsnit i karakter i VØ I og VØ II samlet. Tilsvarende er gældende for køn, hvor kvinder, som tidligere fundet, klarer sig bedre end mænd. Dette skal imidlertid anskues i forhold til den gymnasiale uddannelse, hvor alle koefficienterne er negative. Den samlede effekt er således negativ. Dette passer fint med, at karakteren i VØ I og VØ II som hovedregel

er lavere end såvel det adgangsgivende karaktergennemsnit som den gymnasiale karakter i matematik.

Studenterne med en HF eksamen klarer sig dårligt fulgt af HHX og STX. Dette stemmer også overens med de tidligere fundne resultater. Koefficienten til HTX er ikke signifikant negativ. Det må betyde, at disse studenter klarer sig godt til eksamen i VØ I og VØ II. Dette kan faktisk bekræftes at tabellerne 7 og 8.

Som et supplement er der foretaget en simpel regression, hvor det adgangsgivende karaktergennemsnit er den eneste forklarende variabel til resultatet i VØ I og VØ II. Resultatet af denne regression er vist i tabel 13.

Af tabellen fremgår det, når der sammenlignes med tabel 12, at adgangskvotienten er en meget kraftig forklaringsfaktor for karaktererne i VØ I og VØ II.

Tabel 13: Simpel regression af den gennemsnitlige karakter i VØ I og II

Regressions statistik	
Numerisk korrelation	0.49
R kvadrat	0.24
Justeret R kvadrat	0.24
Standardfejl	2.82
Observationer	1047

ANAVA

	fg	*SS*	*MS*	*F-test*	*P-værdi*
Regression	1	2590.50	2590.50	326.40	0.00
Residual	1045	8293.83	7.94		
Total	1046	10884.34			

	Koefficient	*Std afv.*	*t Stat*	*P-værdi*	*Lav 95 %*	*Høj 95 %*
Skæring	-1.76	0.30	-5.86	0.00	-2.35	-1.17
ADGKVO	0.82	0.05	18.07	0.00	0.73	0.91

7. Afsluttende overvejelser

Undersøgelsen viser, at der er en sammenhæng mellem den gymnasiale karakter i matematik og den opnåede karakter i mikroøkonomi (VØ I og VØ II) på HA/BA-uddannelsen ved SDU. Studenterne klarer sig nemlig markant dårligere i VØ I og VØ II end de resultater, som studenterne har opnået i matematik på de gymnasiale

uddannelser. Dette gælder også, når der ses på det adgangsgivende karaktergennemsnit til HA/BA studiet.

Faldet i karakter er betydeligt, og større i VØ II end i VØ I. Den samlede adgangsgivende karakter i matematik er på 6.9, mens det samlede karaktergennemsnit er på 6.3. Middelkarakterne i VØ I og VØ II er på henholdsvis 2.9 og 2.5, så faldene er på henholdsvis 4.0 og 4.4 point i forhold til karakteren i matematik. Det er signifikant.

Specielt i forhold til VØ I er faldet bemærkelsesværdigt, da studenterne sideløbende med VØ I følger kursus i matematik, hvor der gennemgås eksempler relateret til især emnekredsene i VØ I jævnfør fagbeskrivelsen for kurset, som kan findes på Internettet.

Studenterne har forskellige karakterer med sig fra de gymnasiale uddannelser, og denne karakter har ligeledes en relation til de opnåede resultater. Studenter med HHX, STX og specielt HF klarer sig i VØ I og VØ II signifikant dårligere, end såvel matematikkarakteren som den adgangsgivende karakter tilsiger. Omvendt klarer de internationale studenter og studenter med en uoplyst baggrund sig bedre end forventet. Dette kan for de internationale studenters vedkommende måske forklares med, at studenterne kommer fra en anden uddannelseskultur. Studenter med en baggrund som HTX klarer sig også bedre end forventet. Dette kan måske henføres til denne uddannelses rimelige niveau i matematik.

For HHX, STX og HF er det vanskeligt at pege på årsagsforklaringer. HHX har en handelsvidenskabelig baggrund, og kunne måske tænkes at være interesserede i andre erhvervsøkonomiske discipliner end VØ I og VØ II.

Studenter, der har matematikkarakteren 2 med fra gymnasiet, klarer sig relativt godt i forhold til bestået af VØ I og VØ II i forhold til studenter, der kar matematikkarakteren 12 med sig fra gymnasiet. De mest symmetriske fordelinger af karakterne observeres ved matematikkaraktererne 7 og 10. På denne baggrund kan et højt adgangskrav i matematik ikke siges automatisk, at give en god karakter i VØ I og VØ II. Et højt matematisk karakterkrav kan således ikke anbefales med udgangspunkt i nærværende statistiske materiale.

Modelleres den gennemsnitlige karakter i VØ I og VØ II i forhold til den adgangsgivende karakter, køn og gymnasial uddannelse genfindes mange af de allerede nævnte observationer. Dog viser det sig, at det er den adgangsgivende karakter uanset andre parametre, der er den primære faktor i relation til den opnåede karakter i VØ I og VØ II.

Fear of the unknown – student reluctance to engage in blended learning in undergraduate economics

Julie Buhl-Wiggers, Copenhagen Business School
Annemette Kjærgaard, Copenhagen Business School
Lisbeth la Cour, Copenhagen Business School

Abstract

Blended learning has become one of the most popular approaches to innovate teaching practices in higher education. Yet, few randomized evaluations of the effectiveness of this approach exists. In this study we conduct a randomized evaluation of a blended learning initiative implemented in a large undergraduate macroeconomics course. In particular, the blended learning initiative consist of: 1) an adaptive learning platform, which provides instant feedback on exercises and 2) a new format for the exercises, which facilitates active learning through increased group work. The effects of blended learning are measured by differences in academic achievement. On average, we find no significant differences in the final grades. Yet, when exploring the mechanisms behind this result we find that when adjusting for differences in attendance the treatment increase test scores by 0.15 standard deviations. Our results point to the potential for blended learning to increase academic achievement. Yet, also highlight challenges of getting students onboard and thus point to the importance of academic socialization of students into becoming independent learners as an important part of the success of blended learning.

Extended summary

Background: In recent years, the use of technology in education has increased dramatically and blended learning has become one of the most popular approaches to innovate teaching practices in higher education. Pedagogically there are several reasons why blended learning might increase learning compared to both face-to-face and online learning. First, blended learning increases flexibility, which enables students to have more control over time, place, path or pace of learning. Second, blended learning provides a space for interacting with peers and instructors, which can facilitate increased discussion, collaborative analysis and reflection (Bonk & Graham 2006, Garrison, D. R. & Vaughan, N. D., 2012 . Moreover, blended learning might be a cost-effective way for higher education institutions to tackle the pressure from increasing number of students but decreasing per student revenues. Despite these potentials, few randomized evaluations of the effectiveness of this approach exists (Means *et al.* 2013). In this study we address this gap in the literature by conducting a randomized evaluation of a larger blended learning initiative at Copenhagen Business School to measure the effect of the initiative. We then combine the results with surveys of students learning approach and qualitative interviews about their perception of the initiative in a mixed methods approach.

Hypothesis: The specific problem that we address is that exercise classes in most economics courses are highly teacher-centered with very limited student participation. Our hypothesis is that students will benefit academically from engaging in active learning rather than being passive receivers of teachers' instructions.

Methods: We conduct a Randomized Controlled Trial (RCT) to measure the impact of a new blended learning initiative implemented in a large undergraduate macroeconomics course (N=499). In particular, the blended learning initiative consist of: 1) an adaptive learning platform, which provides instant feedback on exercises and 2) a new format for the exercises, which facilitates active learning through increased learning-by-doing and group work. Hence, we randomly divided the students into a treatment group (blended learning) and a control group (business-as-usual).
The effects of blended learning are measured by differences in academic performance on the final written exam. In addition, we also look at heterogeneous effects along several dimensions, such as gender, personality traits and cognitive ability.
A survey of students' learning preferences and qualitative interviews of students' perceptions of the blended learning initiative complement the RCT to better understand student behavior.

Findings: Looking at the raw averages between the treatment and the control group we find no significant differences in the final grades. This result remains when controlling for teacher specific heterogeneities. Exploring the mechanisms behind this result show that the students in the treatment group were less likely to attend the exercise class compared to students in the control group. Adjusting for this difference in attendance, we find that being in the treatment group increase test scores by 0.15 standard deviations. Turning to heterogeneous effects, we find that more introvert students and students with lower conscientiousness (measured by NEO-FFI-3) are benefitting more compared to more extrovert students and students with higher conscientiousness. We find no heterogeneous effects of gender, cognitive ability, age or other personality traits.

The interviews with students and the observations from the teachers in the initiative show that students are reluctant to engage in the blended learning initiative as they are afraid of letting go of the instructor-directed approach and feel insecure in a more student-directed approach.

Interpretation: These findings point to the potential for blended learning to increase academic achievement, through more effective use of time spend in class. However, our results also highlight challenges of getting the students onboard and engaged as a less teacher-centered approach requires more responsibility from the student, and at first glance might feel like "they are missing out" compared to the traditional approach. Academic socialization of students into becoming independent learners include un-learning of a more passive student behavior established in primary and secondary education. This challenge includes the introduction of a pedagogical approach that actively engages students in developing a more independent learning approach which might be especially important in a higher education context where the students have already formed their study approach and will be an important part of the next phase of this project.

References

Bonk, C. J. & Graham, C. R. (2006). *The Handbook of Blended Learning: Global Perspectives, Local Designs*. Pfeiffer.

Garrison, D. R. & Vaughan, N. D. (2012). *Blended Learning in Higher Education: Framework, Principles, and Guidelines*. Wiley.

Means, B., Toyama, Y., Murphy, R., Baki, M., 2013. The Effectiveness of Online and Blended Learning: A Meta-Analysis of the Empirical Literature. Teach. Coll. Rec. 115.

Erfaringer med eksperimenter der genererer data til brug i undervisningen

Helle Sørensen (helle@math.ku.dk)

Institut for Matematiske Fag, Københavns Universitet

Resumé Jeg har de sidste to år udført små eksperimenter på mine kurser og brugt data som cases i undervisningen. I det ene forsøg løser de studerende sudokoer på tid; i det andet gætter de studerende på antallet af punkter i tre punktplot. Jeg har udført begge eksperimenter på to introducerende kurser og har svar fra 100-200 studerende fra hvert kursus. Data fra eksperimenterne har vist sig at være en guldgrube i undervisningen: De giver anledning til at diskutere begreber som randomisering, uafhængighed, transformation, variansanalyse, reproducerbarhed, og de kan bruges i både forelæsninger og opgaver. Og allervigtigst: De studerende synes det er sjovt og kan nemt forholde sig til problemstillinger, data og resultater. Atiklen forklarer hvordan forsøgene er udført og giver eksempler på hvordan jeg har brugt datasættene i undervisningen.

1 Introduktion

Jeg har de sidste år leget med at få studerende til at deltage i små eksperimenter og derved genereret data som kan bruges i undervisningen. Et af formålene er naturligvis at frembringe nogle datasæt som de studerende har „et forhold" til, som alle har lige gode forudsætninger for at forstå (på tværs af studieretninger), og som består af andet/mere end højde, vægt, alder osv. Et andet formål er at give de studerende en lillebitte fornemmelse for forsøgsplanlægning, dataindsamling og dataklargøring. Data har det ellers med at falde ned fra himlen i pæn indpakning på mine kurser, men selvom man i dag vitterligt kan hente store mængder af data ned fra *skyen*, så er planlægning og praktisk udførelse af eksperimenter stadig et vigtigt element i de eksperimentelle fag. For det tredje er det sjovt—både for de studerende (siger de, når de bliver spurgt) og for mig.

I denne artikel beskriver jeg to eksperimenter, som jeg har testet på to kurser med mere end 100 studerende. I det ene forsøg har de studerende løst sudokoer på tid, i det andet har de gættet på antal punkter i tre punktplot. Nedenfor beskriver jeg forsøgene og den praktiske udførelse (afsnit 2), giver eksempler på hvordan jeg har benyttet data i undervisningen (afsnit 3) og rapporterer resultater fra nogle analyser af data (afsnit 4).

Det er ikke raketvidenskab, og masser af undervisere, herunder mange af mine kolleger, har anbefalet og gjort lignende ting før mig (Brophy and Hahn, 2014; Gelman and Nolan, 2002; Pfaff and Weinberg, 2009; Cummiskey *et al.*, 2012). Ikke desto mindre har jeg haft langt større større nytte af datasættene end jeg umiddelbart havde forestillet mig.

2 Indsamling af data

Begge eksperimenter er udført på to kurser, og inden jeg beskriver eksperimenterne i detaljer, giver jeg en kort beskrivelse af kurserne.

2.1 Kurserne

Kurset *Statistisk Dataanalyse 1* (StatDat1) er obligatorisk på andet eller tredje år på bacheloruddannelserne Husdyrvidenskab, Jordbrugsøkonomi, Naturressourcer og til og med studieåret 2017/18 også på Biologi-Bioteknologi. Der er 150–200 studerende tilmeldt kurset. Kurset er anvendelsesorienteret og handler mest om analyser baseret på normalfordelingen, til og med multipel lineær regression og tosidet variansanalyse.

Kurset *Sandsynlighedsregning og Statistik* (SS) er obligatorisk på første år på bacheloruddannelserne i Matematik, Matematik-Økonomi og Forsikringsmatematik. Der er cirka 300 studerende tilmeldt kurset. Kurset er matematisk af natur og handler blandt om sandsynlighedsteori for kontinuerte fordelinger. Statistikdelen af kurset handler om en og to stikprøver samt lineær regression.

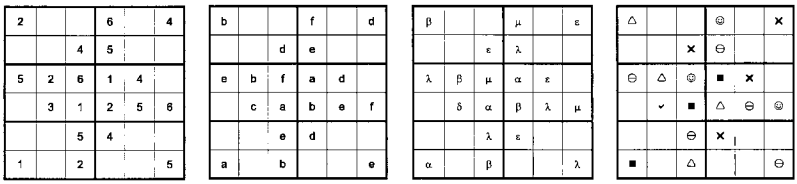
Eksperimentet med punkplot blev udført i studieåret 2017/18, mens sudokoeksperimentet blev lavet i studieåret 2018/19. Begge forsøg blev først lavet på StatDat1.

2.2 Sudoko

Sudokoeksperimentet er stort set replikeret fra Brophy and Hahn (2014). Der er fire forskellige sudokoer, og hver studerende skal løse en af dem på tid. De fire sudokoer har præcis samme struktur, men adskiller sig ved hvilken type tegn der benyttes: tal, latinske

bogstaver, græske bogstaver, eller symboler. De fire sudokoer er vist i Figur 1, gengivet fra Brophy and Hahn (2014).

Sudokoerne blev printet og sorteret så de fire typer lå skiftevis, således at tildelingen af sudokoerne til studerende var randomiseret. Eftersom strukturen af alle sudokoer var ens, fik de studerende instrukser om ikke at kigge hos hinanden eller snakke sammen i løbet af forsøget. Tidtagning blev taget på et full-screen stopur på projektoren.

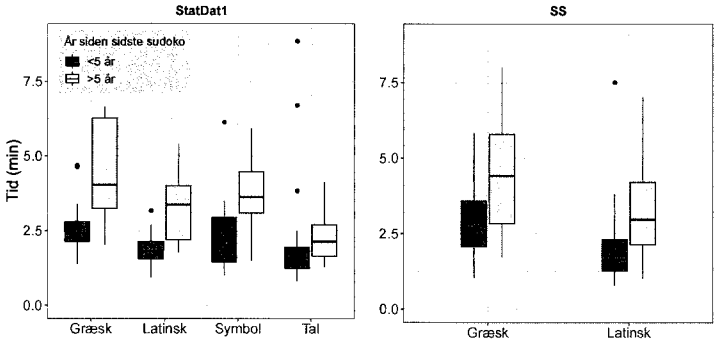


Figur 1: De fire sudokoer. Bemærk at strukturen er ens, altså at den eneste forskel er typen af tegn der indgår.

På StatDat1 udførte jeg eksperimentet ved en forelæsning og lod forsøget vare ti minutter. Dette viste sig at være lang tid (højeste observation var under ni minutter, næsthøjeste under syv minutter), og jeg lod derfor kun eksperimentet køre otte minutter på SS. På SS blev eksperimentet desuden udført af lærerne på kursets ti øvelseshold, og jeg brugte kun sudokoer med latinske og græske bogstaver fordi pensum på dette kursus kun omfatter sammenligning af to stikprøver.

Efter de ti henholdsvis otte minutter blev de korrekte løsninger vist til de studerende, så de kunne checke om de havde løst deres sudoko korrekt. Derefter udfyldte deltagerne en elektronisk formular hvor de skulle indtaste hvor lang tid de havde brugt på at løse sudokoen, om de havde løst den korrekt, og desuden svare på en række spørgsmål vedr. studieretning, erfaring med at løse sudokoer, holdnummer (kun SS) mm. Med hensyn til erfaring blev de studerende blev spurgt hvornår de sidst havde løst en sudoko og kunne svare *mindre end 5 år siden* eller *mere end 5 år siden* eller *aldrig*. Meget få svarede at de aldrig havde løst sudokoer, så de sidste to kategorier er slået sammen. Det tog 15–20 minutter at udføre forsøget, inklusive forklaringer og indtastning.

Figur 2 viser boksplot over løsningtiden, opdelt efter kursus (StatDat1 til venstre og SS til højre) samt sudokotype og erfaring. Data ser ud cirka som man kunne forvente: Sudokoer med tal og latinske bogstaver er hurtigere at løse end de øvrige typer, og det hjælper at have løst sudokoer „for nylig".

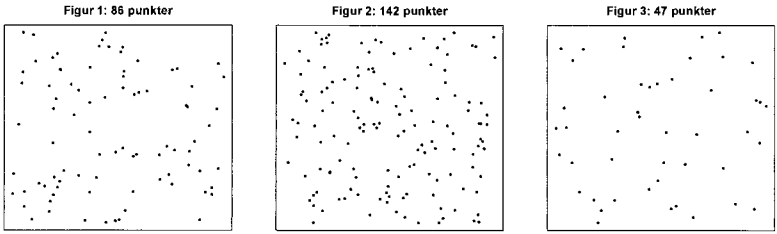


Figur 2: Boksplot over løsningstiden fra sudokoforsøget opdelt efter kursus, sudokotype og erfaring. Der er data fra 112 studerende på StatDat1 og 192 studerende på SS. Der er cirka lige mange studerende for hver sudokotype (indenfor kursus), mens kategorien *<5 år* omfatter cirka 60% af deltagerne.

2.3 Punktplot

Eksperimentet med punktplot er inspireret af data indsamlet af min tidligere kollega Claus Ekstrøm (Ekstrøm, 2016, afsnit 3.39). Det blev udført ved forelæsninger på begge kurser. Jeg viste tre punktplot på projektoren; et ad gangen, hvert plot i cirka seks sekunder og med en pause mellem plottene. De studerende gættede på antallet af punkter i hvert plot undervejs og angav umiddelbart efter forsøget deres gæt ved at udfylde en elektronisk formular hvor de også svarede på diverse spørgsmål vedr. køn, studieretning mm. De studerende fik strenge instrukser om ikke at snakke sammen undervejs. De sande antal blev først afsløret for de studerende senere i kurset. Det tog under 10 minutter at udføre forsøget, incklusive forklaringer og indtastning af data.

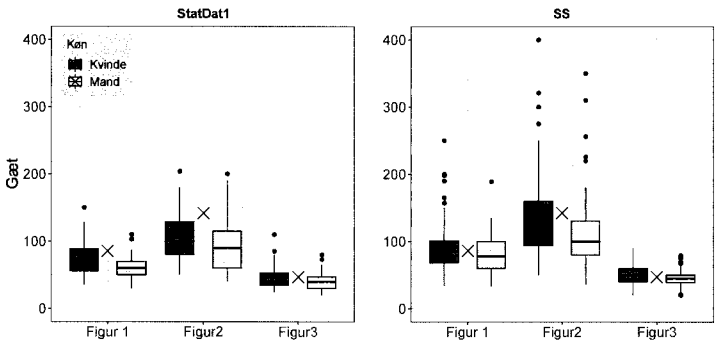
Punktplottene er vist i figur 3. Det sande antal punkter er angivet over punkplottene; det var natuligvis ikke tilfældet da de blev vist til de studerende. Figur 4 viser boksplot over de studerendes gæt, opdelt efter kursus (StatDat1 til venstre og SS til højre) samt figur og køn. Krydserne repræsenterer det sande antal punkter i figurerne. Særligt for figur 2 med mange punkter ses det at det typiske gæt er for lavt.



Figur 3: De tre punktplot som blev vist til de studerende. Det sande antal punkter er angivet øverst; dette blev ikke vist til de studerende.

3 Brug af data i undervisningen

Jeg har brugt datasættene som gennemgående cases i kurserne; ved flere forelæsninger og til adskillige øvelsesopgaver og/eller afleveringsopgaver. Sidste år brugte jeg desuden data fra SS til reeksamen på StatDat1, og jeg har desuden lånt data til flere kolleger som har brugt dem til eksamensopgaver på andre kurser.



Figur 4: Boksplot over gættene på antal punkter opdelt efter kursus, figur og køn. Krydserne viser det sande antal punkter i de tre figurer (86, 142, 47). Der er data fra 140 studerende på StatDat1 og 191 studerende på SS.

Oprindelig var tanken først og fremmest at bruge data til illustration af variansanalysemodeller, men data vist sig nyttige til at diskutere en lang række problemstillinger og emner—flere end jeg havde forestillet mig inden jeg havde data. Lad mig give nogle eksempler.

Randomisering, konfundering I forbindelse med sudokoeksperimentet diskuterer jeg vigtigheden af at arkene med de fire sudokotyper ligger skiftevis ved uddelingen i lokalet. På StatDat1 er der en tendens til at studerende fra samme studieretning også sidder sammen i lokalet, og de studerende kan nemt forholde sig til konfunderingsproblemer hvis sudokoerne i stedet havde været sorteret så alle af samme type lå sammen. Tilsvarende diskuterer jeg vigtigheden af at de fire sudokoer er helt ens bortset fra typen af tegn, således at en eventuel forskel mellem typerne ikke kan skyldes forskellig sværhedsgrad. Igen noget som de studerende nemt kan forholde sig til—faktisk var der en studerende der bragte det på bane allerede inden jeg var nået til det.

Uafhængighed Som nævnt i afsnit 2.2 gjorde jeg i begge eksperimenter et stort nummer ud af at de studerende ikke må snakke sammen i løbet af forsøget. Om det virker ved jeg ikke, og i virkeligheden er det væsentlige er at de studerende kan huske instruksen når vi senere snakker om antagelsen om uafhængighed mellem data fra forskellige personer. Omvendt kan simple scatterplots og korrelationer mellem gættene på de tre figurer bruges til at illustrere at data fra samme person bestemt ikke er uafhængige.

Dataklargøring, fejl i data, manglende data De studerende indtastede deres løsningstider/gæt samt baggrundsoplysninger på en elektronisk formular. Adskillige ting kan gå galt i den forbindelse. For eksempel blev de i sudokoforsøget bedt om at indtaste et helt antal minutter samt antal sekunder udover de hele minutter. Hvad betyder det så når en studerende indtaster 1 min og 96 sek? En anden studerende angav tiden 9 min og 46 sek på trods af at forsøget blev stoppet efter 8 min. Første gang jeg kørte punktploteksperimentet, var jeg tilsyneladende ikke helt skarp i mine instrukser; i hvert fald indtastede cirka 10 sturende deres gæt allerede inden jeg havde vist figurerne på projektoren! Disse observationer skulle oplagt fjernes fra datasættet, men hvad gør man med andre gæt på fx 10 eller 1000? Det er næppe seriøse gæt, men hvor går grænsen? Derudover var der studerende der ikke svarede eller ikke svarede entydigt på alle baggrundsspørgsmål; kan man alligevel bruge data fra disse personer?

Log-transformation Den problemstilling som datasættene måske har belyst allerbedst, er transformation, herunder fortolkning af estimater når analysen er kørt på log-transformerede data.

Gættene på antal punkter viser sig på smukkeste vis at være log-normalfordelte; der er nogle der gætter alt, alt for højt, men få der gætter meget lavt. QQ-plots for figur 1 fra

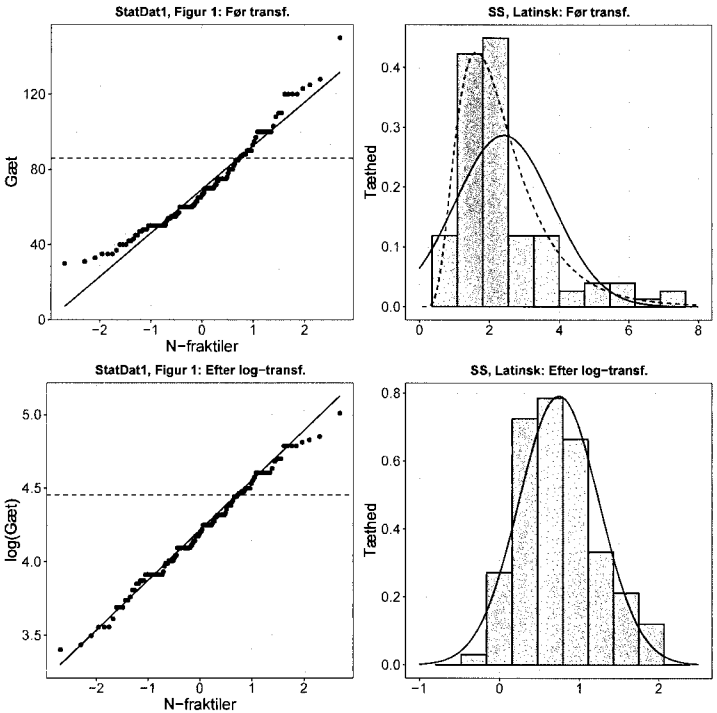
StatDat1 er vist i venstre side af figur 5 før og efter log-transformation. På StatDat1 er der mange datasæt der kræver log-transformation, og jeg tror næsten at de studerende antager at det er noget jeg med vilje opsøger (for at genere dem), så det er glimrende at de på egen hånd frembringer data der lægger op til transformation. Log-transformationen fører naturligt til en diskussion af hvordan det tilbagetransformerede estimat skal fortolkes som en estimeret median, og hvordan fx forskellen mellem mænd og kvinder skal fortolkes som en relativ, snarere end en absolut, forskel. Det er også oplagt at diskutere variationskoefficienter: På StatDat1 var de næsten ens for de tre figurer (0.329, 0.336 hhv. 0.319 når der ikke justeres for andre variable).

Størstedelen af SS handler om sandsynlighedsregning, og det er først sent i kurset at der er egentlige statistiske analyser, men transformationsproblematikken fungerer udmærket alligevel. En del af den obligatoriske opgave i 2018/19 bestod i at tegne histogrammer og normalfordelingstætheder for løsningstiden for sudokoer med latinske bogstaver, før og efter transformation; at udlede tætheden for log-normalfordelingen og sammenligne med data; samt bestemme middelværdi og median for log-normalfordelingen. Histogrammer og tætheder er vist i højre side af figur 5.

Lineære normale modeller Data fra punktplotforsøget fungerer glimrende som eksempel på en enkelt stikprøve fordi det sande antal punkter kan bruges som referenceeller hypoteseværdi (og ikke er nul). Det viser sig at være særligt sjovt fordi det åbenbart er svært at gætte medmindre der kun er få punkter i plottet. Et lille aber dabei: På dette stadie af undervisningen taler jeg gerne om populationer og stikprøver, men der er ikke for alvor en naturlig population hvorfra vi har udtaget en repræsentativ stikprøve.

For begge eksperimenter ligger det lige til højrebenet at køre variansanalyser (på log-transformerede data). På StatDat1 gør jeg det ad flere omgange som kurset skrider frem, dvs. inddrager først en enkelt forklarende variabel (fx sudokotype), derefter flere variable og enkelte vekselvirkninger. Vekselvirkninger er notorisk svært, men bliver mere tilgængeligt i denne indpakning: Er forskellene mellem sudokotyperne fx forskellig for de forskellige niveauer af erfaring? På SS er ambitionsniveauet mindre da pensum kun omfatter en og to stikprøver.

Data er mindre velegnede til regressionsanalyser, men en (måske lidt søgt) opgave om multipel regression på StatDat1 viste dog at gæt på figur 1 og/eller 3 kan inddrages til prædiktion af en persons gæt på antal punkter i figur 2, og man kunne have diskuteret *regression towards the mean* i samme forbindelse.



Figur 5: Illustrationer vedrørende log-transformation. Til venstre ses QQ-plots for gæt på antal punkter i figur 1 før/efter (øverst/nederst) log-transformation. Til højre ses histogrammer for løsningstider for sudokoen med latinske bogstaver før/ efter log-transformation (øverst/nederst). De ikke-stiplede kurver til højre er normalfordelingstætheder, mens den stiplede kurve i øverste højre graf er tætheden for den estimerede log-normalfordeling.

Binomialmodeller, logistisk regression Oprindelig var det min plan at korrekthed af sudokoen kunne bruges som en binær respons men det viser sig at næsten alle løser sudokoerne korrekt. Til gengæld kan data fra punktplotforsøget bruges binært hvis man udelukkende fokuserer på om gættene er for høje eller lave.

Reproducerbarhed I punktplotforsøget på StatDat1 viste der sig at være signifikant kønsforskel; det estimeres at kvinder gætter bedre end mænd (se venstre side af tabel 2). Det havde de (kvindelige) studerende en del morskab af, og det førte naturligt til en diskussion om at man altid bør gentage eksperimenter der har ført til overraskende konklusioner, og derved (forsøge at) konfirmere resultaterne. I dette tilfælde *blev* forsøget jo faktisk gentaget på SS, og med samme konklusion (se højre side af tabel 2), så de kvindelige SS-studerende kunne hovere dobbelt så meget.

Hvis det hørte naturligt hjemme på kurserne, kunne data også illustrere andre temaer. Data fra sudokoeksperimentet kunne bruges i forbindelse med overlevelsesanalyse, se evt. Kaplan-Maier plots i Brophy and Hahn (2014), og man kunne afkorte forsøget hvis man gerne vil diskutere censorering i højere grad. Sudokoeksperimentet blev udført på øvelsesholdene på SS, hvilket naturligt lægger op til at inddrage en tilfældig holdeffekt, ligesom en tilfældig personeffekt bør inddrages hvis man analyserer gæt på flere figurer i en samlet model. Der er givetvis andre muligheder som jeg ikke har tænkt over.

4 Resultater

Resultaterne fra de statistiske analyser er selvsagt ikke det vigtigste for den didaktiske diskussion. Og dog: Det er ikke uvæsentligt for succesoplevelsen at resultaterne er forståelige og måske ligefrem efterviser de teorier/hypoteser man måtte have på forhånd. På kurser med studerende fra eksperimentelle fag er dette en pointe i sig selv. Data er analyseret separat for de to kurser.

4.1 Sudoko

For sudokoeksperimentet har jeg brugt en tresidet variansanalyse uden vekselvirkninger med logaritmen til løsningstiden som respons og følgende forklarende variable: Sudokotype (fire typer for StatDat1, to typer for SS, jf. figur 1); studieretning (fire niveauer for hvert kursus), samt erfaring (mere eller mindre end fem år siden sidst løste sudoko).

Tabel 1 viser p-værdierne for effekten af hver variabel samt estimater for effekter der kan estimeres på begge kurser, se også figur 1. Estimaterne er tilbagetransformerede og angiver dermed den faktor som estimerede medianer vokser med når den givne faktor ændres. Estimaterne er tæt på ens for de to kurser: Mediansløsningstiden estimeres til at være 45% hhv. 47% højere for sudokoer med græske bogstaver sammenlignet med latinske bokstaver, og medianløsningstiden estimeres 60% hhv. 64% højere for studerende der ikke har løst sudokoer inenfor de sidste fem år sammenlignet med studerende der faktisk har løst sudokoer indenfor de sidste fem år. På StatDat hvor fire sudokotyper blev testet, var løsningstiden lavest for sudokoer med tal og 64%, 13% hhv. 33% højere for sudokoer med græske bogstaver, latinske bogstaver hhv. symboler.

	StatDat1	SS
Antal studerende	112	192
Test for hovedeffekter		
Sudokotype	$p = 0.00039$	$p = 1.6 \cdot 10^{-9}$
Erfaring	$p = 1.4 \cdot 10^{-7}$	$p = 5.2 \cdot 10^{-13}$
Studieretning	$p = 0.017$	$p = 0.024$
Estimat (95% KI) for faktor		
Græsk vs. Latinsk	1.45 (1.15 – 1.84)	1.47 (1.30 – 1.66)
> 5 år vs. < 5 år siden sidst	1.60 (1.36 – 1.89)	1.64 (1.45 – 1.86)
Variationskoefficient	0.441	0.440

Tabel 1: Resultater fra sudokoeksperimentet: p-værdier og udvalgte estimater fra en tresidet variansanalyse for hvert kursus. Estimater er tilbagetransformerede, så de angiver den estimerede faktor som medianen forøges med, når variablen ændres.

4.2 Punktplot

Gættene på antal punkter i figur 1 er analyseret med en tosidet variansanalyse uden vekselvirkning med logaritmen til gættet som respons og køn og studieretning (fire niveauer for hvert kursus) som forklarende variable.

Tabel 2 viser udvalgte resultater, se også figur 3. Mænd estimeres til at gætte 22% hhv. 17% højere end kvinder, og forskellen er statistisk signifikant på begge kurser. Der er 86 punkter i figur 1, og denne værdi er indeholdt i konfidensintervallet for kvinder på SS (gennemsnit over studier), men hverken for mænd på SS eller for kvinder og mænd på

StatDat1.

I de tilsvarende analyser for figur 2 og figur 3 gættede kvinder også signifikant højere end mænd (begge figurer, begge kurser), men analyserne af de tre figurer er oplagt ikke uafhængige så det skal man næppe lægge for meget i. Ikke overraskende gætter man bedre når der er få punkter i figuren: For figur 2 med 142 punkter var det sande antal ikke indeholdt i konfidensintervallet (begge køn, begge kurser), mens det for figur 3 med 47 punkter kun er mænd fra StatDat1 hvor det sande antal ikke er indeholdt i konfidensintervallet.

	StatDat1	SS
Antal studerende	140	191
Test for hovedeffekter		
Køn	$p = 0.0030$	$p = 0.0055$
Studieretning	$p = 0.69$	$p = 0.66$
Estimat (95% KI) for forhold		
Kvinder vs. Mænd	1.22 (1.07 – 1.39)	1.17 (1.05 – 1.30)
Estimeret median (95% KI)		
Kvinder, gennemsnit over studier	70.3 (65.5 –75.3)	88.0 (80.2 – 96.6)
Mænd, gennemsnit over studier	57.5 (51.8 – 63.7)	75.5 (69.8 – 81.6)
Variationskoefficient	0.322	0.387

Tabel 2: Resultater vedr. figur 1 fra punktploteksperimentet: p-værdier og udvalgte estimater fra en tosidet variansanalyse for hvert kursus. Estimatet for effekten af køn er tilbagetransformeret, så det angiver estimatet for forholdet mellem mediangæt for kvinder og mænd. Det sande antal punkter i figur 1 er 86.

5 Konklusion

Datasættene har vist sig at være guld værd i undervisningen. Det er sjovest at bruge dem sammen med de studerende der har været udsat for eksperimenterne, men de fungerer også fint i efterfølgende kurser fordi de er nemme at forholde sig til. På StatDat1 spurgte jeg de studerende direkte om de synes det var sjovt/irriterende/ligegyldigt at inddrage data som de selv havde været med til at frembringe. Cirka 80% svarede at det var sjovt, resten at det var ligegyldigt, og flere nævnte det sjove i den „konkurrence" der fx opstod mellem studieretninger.

Min erfaring er altså at de studerende synes det er sjovt, og også at de er villige til at lege med. Jeg forklarer omhyggeligt at det er frivilligt at deltage men at det er sjovest hvis mange deltager og svarer ærligt og seriøst. Der er som regel et par stykker der svarer fjollet/useriøst, men det er ganske få. Selvom der ikke spørges om personfølsomme ting, skal man selvfølgelig sørge for at enkeltpersoner ikke kan genkendes udfra baggrundsoplysningerne. Og kan man overhovedet spørge til studerendes køn i 2018?

De to eksperimenter kan udvides med flere sudokoer (af forskellig sværhedsgrad), og flere punktplot (gerne også gentagelser af de samme punktplot) således at man får mere komplicerede datastrukturer, men logistikken bliver sværere. Det er en pointe i sig selv at det er praktisk muligt at gennemføre hvert forsøg på 10–20 minutter, også på kurser med mange studerende, og en væsentlig del af succesen stammer fra at (næsten) alle deltager.

Som det forhåbentlig fremgår af artiklen, har jeg selv haft stor fornøjelse af forsøgene. Det er skægt at udføre disse små forsøg og—særligt i sudokoeksperimentet—se hvordan resultaterne i høj grad bliver som man ville forvente.

Litteratur

Brophy, C. and Hahn, L. (2014). Engaging students in a large lecture: An experiment using sudoku puzzles. *Journal of Statistics Education*, **22**.

Cummiskey, K., Kuiper, S., and Sturdivant, R. (2012). Using classroom data to teach students about data cleaning and testing assumptions. *Frontiers in Psychology*, **3**.

Ekstrøm, C. T. (2016). *The R Primer*. CRC Press, second edition.

Gelman, A. and Nolan, D. (2002). *Teaching Statistics: A bag of tricks*. Oxford University Press.

Pfaff, T. J. and Weinberg, A. (2009). Do hands-on activities increase student understanding?: A case study. *Journal of Statistics Education*, **17**.

Against Average

Symposium i Anvendt Statistik, januar 2019

Gorm Gabrielsen, Lektor emeritus på CBS

Politiken meddelte en uges tid før jul i en stor overskrift: *Klima og miljø tager tigerspring til toppen af vælgernes dagsorden.* Dette er udledt fra en vælgerprofil, der angiver hvor mange der peger på *hvilke områder* som vigtigste politikområder. Men vi får ikke noget at vide om hvorvidt det er de enkelte vælgere, der har taget tigerspringet, eller om der er blevet flere vælgere, der har sat *Klima og Miljø* højst på dagsordenen. Det skyldes, at man ganske enkelt har talt op hvor mange, der peger på de enkelte områder. Man har således fået en profil af den "gennemsnitlige" vælger. Men man kan ikke skelne mellem *hvor mange* og *hvor meget.* Netop denne skelnen er måske ikke så interessant hvis man vil forudsige udfaldet af et kommende valg, men på spørgsmålet *hvad skal vi gøre?* - altså på handlingsplanet kan det være vigtigt.

Men hvad kan man gøre andet? Man kan "prøve at" definere en profil på individniveau. En profil på individ-niveau er en latent struktur, som skal estimeres individuelt. Man skal således for det enkelte individ forstå hvilke områder, som er vigtige politikområder og hvordan disse områder relateres til hinanden. Korrelation skal således være på individniveau og ikke på populationsniveau.

Gennem flere cases vil jeg vise,

- hvordan en latent struktur kan være en ratio-skala
- hvordan (med inspiration fra Rasch) en ratio-skala kan estimeres
- hvilke konsekvenser en sådan tilgang har for den efterfølgende beslutningsproces
- hvordan den latente struktur (ratioskalaen) leder til en alternativ tænkning om segmentering

BLAND-ALTMAN-PLOTS FOR FLERE END TO OBSERVATØRER

SÖREN MÖLLER

OPEN - Odense Patient data Explorative Network, Odense Universitetshospital og Klinisk Institut, Syddansk Universitet, Odense og Odense Agreement Working Group

INDLEDNING

I sundhedsvidenskabelige studier er der ofte et ønske om at undersøge, hvorvidt forskellige observatører (personer eller medicinsk udstyr) er i stand til at måle samme egenskab ved individer (patienter, røntgenbilleder, vævsprøver, ...) med god nok overenstemmelse, til at der i efterfølgende analyser og/eller senere studier, kan ses bort fra, hvilken observatør der har foretaget målingen. Typisk vil denne situation blive statistisk analyseret med *mixed effect modeller* eller lignende tilgange (se [4] og [3]). Dog ønsker kliniske forskere ofte også en simplerere grafisk fremstilling af situationen, og den mest udbredte af disse er Bland-Altman-plottet ([1]) som afbilder observerede forskelle mellem to observatører mod den gennemsnitlige observation for hvert individ, sammen med et 95%-prædiktionsbånd (se nederste række i figur 1 for eksempler).

En af ulemperne ved det klassiske Bland-Altman-plot er, at det kun virker for to observatører, mens det både af praktiske årsager og for at øge ekstern validitet ofte er ønskeligt at inkludere flere end to observatører i kliniske studier. Den klasiskse anbefaling ([2]) i disse situationer er at generere parvise Bland-Altman-plots, hvilket dog bliver uoverskueligt og svært at fortolke, når man kommer over tre observatører, da dette giver både et stort antal figurer og mange parvise sammenligninger at vurdere. Andre forslag indebærer at der indtegnes et punkt for hver observation (i stedet for, for hvert individ) som enten forbindes med linjer ([3]) eller markeres med forskellige symboler ([5]), tilgange der kan blive uoverskuelige ved mange observationer. Ønsket er at udvikle en generalisering af Bland-Altman-plottet til flere end to observatører, der kun kræver én figur med ét punkt per observation.

SCENARIET

Statistisk set er vi interesseret i at undersøge en situation med m observatører, der hver observerer n individer én gang, resulterende i $m \cdot n$ observationer i alt. Vi antager at den data-genererende proces for observationerne $y_{i,j}$ kan beskrives med formlen

$$Y_{i,j} = \mu_i + \gamma_j + \epsilon_{i,j} \qquad i = 1..n, j = 1..m$$

hvor μ_i og γ_j er deterministiske og $\epsilon_{i,j} \sim N(0, \sigma_{i,j})$ parvis uafhængige. Desuden antager vi at $\sigma_{i,j}$ kun afhænger af μ_i, men ikke i.

Det klassiske Bland-Altman-plot er tegnet ud fra en hypotese om at $\gamma_j = 0$ for alle j og $\sigma_{i,j} = \sigma$ for alle (i, j) og det kan derved bruges til at detektere afvigelser fra denne hypotese.

KLASSISK BLAND-ALTMAN-PLOT ($m = 2$)

I det klassiske Bland-Altman-plot (eksempler i nedre række i figur 1) bestemmes punkterne

$$(\bar{x}_i, d_i) := \left(\frac{y_{i,1} + y_{i,2}}{2}, y_{i,1} - y_{i,2} \right)$$

Tak til Birgit Debrabant (SDU), Ulrich Halekoh (SDU) og Oke Gerke (OUH) for samarbejde om projektet i *Odense Agreement Working Group* og tak til Karen Steenvinkel Pedersen (UCL) for gode kommentarer til manuskriptet.

og disse indtegnes i et plot sammen med en vandret middelværdislinje ved

$$\bar{d} := \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} x_{i,1} - \frac{1}{n}\sum_{i=1}^{n} x_{i,2}$$

og to grænselinjer ved

$$(U, L) := \bar{d} \pm t_{0{,}975,n-1} \cdot \sqrt{1 + \frac{1}{n}} \cdot s_d$$

hvor

$$s_d := \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2}$$

er spredningen af de observerede differencer. Bemærk at disse grænselinjer svarer til et 95%-prædiktionsinterval for en ny, normalfordelt difference.

I nogle kilder bliver der i stedet for $t_{0{,}975,n-1} \cdot \sqrt{1 + \frac{1}{n}}$ brugt aproksimative grænser, typisk enten $1{,}96$ svarende til grænsen for store n eller 2 for t-kvantilen og delvis med $\sqrt{1 + \frac{1}{n}}$ udeladt. Vi vil i det følgende antage at $t_{0{,}975,n-1} \cdot \sqrt{1 + \frac{1}{n}}$ bruges som koefficient for grænselinjerne.

Generaliseret Bland-Altman-plot ($m \geq 2$)

Formålet med dette projekt er at forelå en afart af Bland-Altman-plottet, der kan anvendes på flere end to observatører. Specielt ønskes at:

- Der stadig kun er ét punkt på plottet per individ (modsat nogle konkurrerende foreslag i litteraturen, med et punkt per observation, hvilket hurtigt bliver uoverskueligt).
- x-aksen i plottet stadig giver udtryk for den absolutte størrelse af individets måling og y-aksen for variationen mellem observatører på dette individ.

Inspireret af *mean variance plot* som f.eks. bruges i modeltjek for linjær regression foreslår vi at afbilde

$$(\bar{x}_i, s_i) := \left(\frac{1}{m}\sum_{j=1}^{m} y_{i,j}, \sqrt{\frac{1}{m-1}\sum_{j=1}^{m} (x_{i,j} - \bar{x}_i)^2} \right).$$

Dette svarer i $m = 2$ situationen til Bland-Altman-plottet *foldet* langs middelværdislinjen og skaleret med en faktor $\sqrt{2}$ på y-aksen (for at gå fra differencer til spredninger). Da $s_i \geq 0$ vil plottet kun ligge i den positive halvplan og middelværdislinjen erstattes af x-aksen. Eksempler for $m = 2$ situationen kan ses i øvre række i figur 1 og for forskellige valg af m og n i figur 2.

Hvor skal linjen ligge? Mens det er relativt oplagt hvor punkterne skal ligge i ovenstående konstruktion, mangler vi stadig at fastlægge hvor en linje der erstatter 95%-prædiktionsbåndet fra det klassiske Bland-Altman-plot skal ligge.

Vi husker at den observerede varians af k uafhængige normalfordelte observationer med ens middelværdi vil være χ^2-fordelt med $k-1$ frihedsgrader og at spredningen derfor vil være χ-fordelt med $k-1$ frihedsgrader. Skalerer vi dette til vores data finder vi asymptotisk

$$\frac{s_i}{\sigma}\sqrt{k-1} \sim \chi(k-1).$$

Vi vælger derfor at placere vores linje ved

$$L = q_{\chi(m-1)}(0{,}95)\frac{1}{\sqrt{(m-1)}}s \tag{1}$$

hvor $q_{\chi(m-1)}$ betegner kvantilfunktionen for en χ-fordeling med $m-1$ frihedsgrader og $s = n^{-1}\sum_i^n s_i$ er den gennemsnitlige intra-individ-spredning. Vi bemærker at denne formel kun tager højde for antal

observatører, men hverken antal observationer (dvs. usikkerheden i s, som i det klassiske Bland-Altman-plot modelleres ved brug af t-fordelingen) eller at linjen skal være svare til et prædiktionsinterval (faktoren $\sqrt{1+n^{-1}}$ i formlen for det klassiske Bland-Altman-plot), derfor må vores formel forventes kun at være korrekt for store n.

Simulation. For at undersøge, hvor retvisende vores valg af L er, simulerede vi 1000 gange for $m = 2, 3, 4, 5$ observatører og $n = 10, 20, 100$ observationer fra normalfordelinger svarende til ovenstående datagenererende proces og afbilleder den empiriske 95%-kvartil (både for de oprindelige observationer og for 5000 nye observationer) sammen med L bestemt ud fra formel 1:

		Oprindelige observationer			5000 nye observationer		
Individer		n=10	n=20	n=100	n=10	n=20	n=100
Observatører	formel	95% kvartil	95% kvartil	95% kvartil	95% kvartil	95% kvartil	95% kvartil
m=2	1,959964	1,912083	1,927279	1,956579	2,223729	2,075538	1,981387
m=3	1,730818	1,687289	1,707918	1,719955	1,853011	1,80814	1,739351
m=4	1,613973	1,582155	1,595075	1,612125	1,725425	1,661864	1,623649
m=5	1,540108	1,50826	1,518143	1,534751	1,618284	1,579867	1,546908

Tilsvarende bestemmer vi i simulationerne hvor stor en del af observationerne (henholdsvis de oprindelige og 5000 nye) der i gennemsnit ligger under linjen L:

	Oprindelige observationer			5000 nye observationer		
Individer	n=10	n=20	n=100	n=10	n=20	n=100
Observatører	Overdækning	Overdækning	Overdækning	Overdækning	Overdækning	Overdækning
m=2	0,9562	0,9543	0,9505	0,9219	0,9371	0,9475
m=3	0,9598	0,9539	0,9517	0,9300	0,9369	0,9485
m=4	0,9583	0,9541	0,9504	0,9450	0,9397	0,9479
m=5	0,9591	0,9555	0,9514	0,9303	0,9399	0,9483

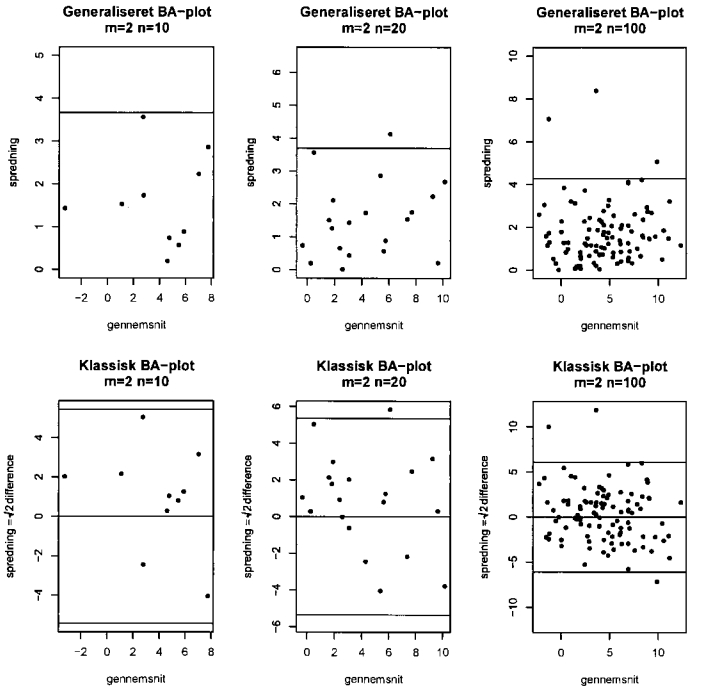
Diskussion

Ud fra simulationen kan vi se at vores linje generelt resulterer i en overdækning omkring 93% ved $n = 10$ observationer stigende til omkring 94,8% ved $n = 100$ observationer, hvilket forventes at være tilfredsstillende til de fleste praktiske formål. Vi observerer dog også at den empiriske 95%-kvartil af nye observationer generelt ligger et stykke over L, specielt for små n, så en mere præcis formel for L ville være ønskeligt.

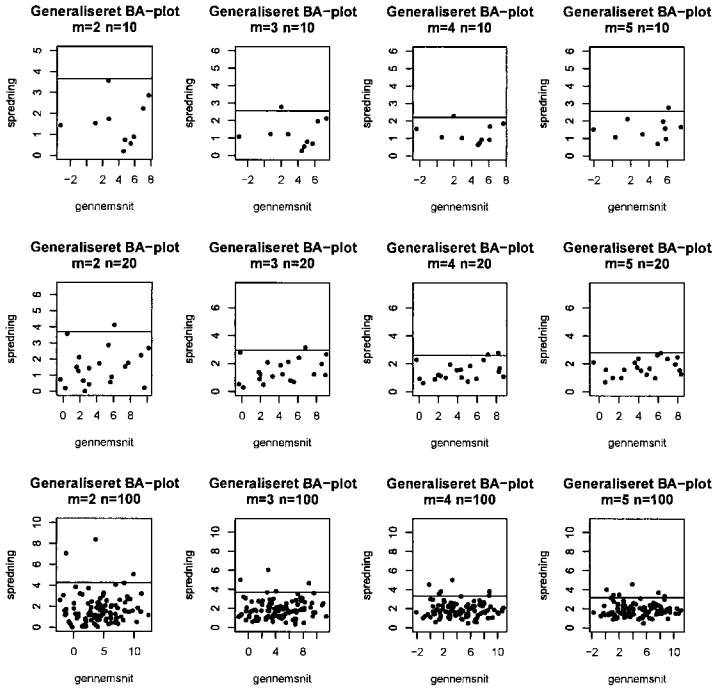
Et af formålene med det klassiske Bland-Altman-plot er at kunne detektere om variansen vokser med voksende individgennemsnit, dette er simuleret i figur 3 for både det klassiske og vores generaliserede Bland-Altman-plot og det indikerer at sådanne problemer er detekterbare på vores plot.

En af udfordringerne med vores forslag er at det ikke er tydeligt om nogle af obersvatørerne udviser et systematisk bias, svarende til en forskubning af midterlinjen i det klassiske Bland-Altman-plot. Desuden er det ikke tydelig fra plottet, hvilken oberservatør der er skyld i stor variation af et individs observationer. Vi håber fremadrettet at kunne udvide vores plot med indikationer af disse faktorer.

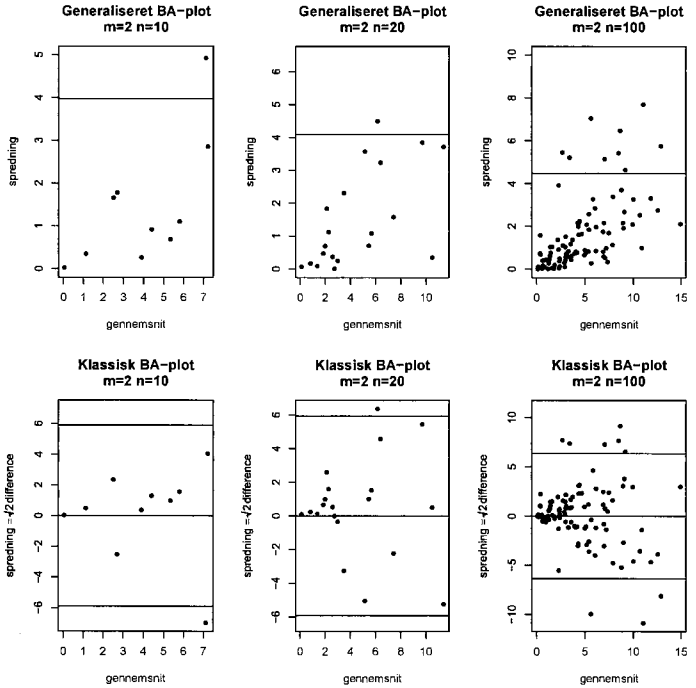
FIGUR 1. Sammenligning mellem generaliseret og klassisk BA-plot for to oberservatører.



FIGUR 2. Generaliseret BA-plot for forskellige antal observatører og observationer.



FIGUR 3. Sammenligning mellem generaliseret og klassisk BA-plot for to observatører ved spredning voksende med voksende gennemsnit.



REFERENCER

[1] J. M. Bland and D. G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement*, Lancet **327** (1986), no. 8476, 307–310.

[2] ———, *Measuring agreement in method comparison studies*, Stat Methods Med Res **8** (1999), no. 2, 135–160.

[3] B. Carstensen, *Comparing clinical measurement methods*, Wiley, 2010.

[4] O. Gerke, S. Möller, B. Debrabant, and U. Halekoh, *Experience from applying the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) indicated 5 questions to be addressed in the planning phase from a statistical point of view*, Diagnostics **8** (2018), no. 4.

[5] M. Jones, A. Dobson, and S. O'Brian, *A graphical method for assessing agreement with the mean between multiple observers using continuous measures*, International Journal of Epidemiology **40** (2011), 1308–1313.

Statistiske analyser af vægtede data

- repræsentativitet, teori og praksis

Peter Linde, Det Nationale Forskningscenter for Arbejdsmiljø

Statistiske test og repræsentativitet – helt enkelt

En statistisk t-test af en hypotese H: $\mu_1 = \mu_2$ er i sin simple udgave ret enkel

$$teststørrelse = \frac{\hat{x}_1 - \hat{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \qquad (1)$$

I tælleren er differensen på de to estimater for vores parametre μ_1 og μ_2, der knytter sig til hypotesen, og i nævneren spredningen af variansen for disse to estimater. Hvor tættere vores teststørrelse er 0, hvor mere sandsynlig er hypotesen. Da estimaterne er baseret på en stikprøve falder varianserne proportionalt med stikprøverne. Jo større stikprøve, jo mindre bliver vores varians under kvadratroden.

Der er selvfølgelig en række forudsætninger for, at denne test giver mening.

Den første, og på en måde den vigtigste, er at $E(\hat{x}_i) = \mu_i$. (2)

Dette grundlæggende krav betyder, at vores estimater er middelrette, uden bias eller centrale – tre udtryk for det samme. Ellers giver test-størrelsen ikke mening. Hvis det er en befolkningsundersøgelse, og vi udtaler os om forhold i denne population, siger vi, vi har repræsentative estimater, hvis der ikke er nogen bias. Hvis der er en bias, vil denne bias komme oven i forskellen mellem de to estimater, dvs. i tælleren i formel (1). Og hvis stikprøven er tilstrækkelig stor, vil denne bias give en signifikant test, selvom hypotesen H er opfyldt. Så hvis vi ikke har repræsentative estimater, er vi på et tidspunkt sikker på at få signifikante test, hvis vi øger stikprøven, selv om der ingen forskel er.

I en dataindsamling kommer man i praksis til at leve med en vis bias. Det er endnu ikke lykkedes at lave den perfekte dataindsamling uden fejl. Svaret herpå er, som antydet overfor, ikke først og fremmest at øge stikprøven. Det eneste vi er sikre på, når vi ikke har repræsentative estimater, er at vi med en stor stikprøve er endnu mere

sikker på, at estimaterne er forkerte. For det er sikkerhedsintervallet omkring de forkerte estimater, vi reducerer, når vi øger stikprøven. Det er ikke estimaternes bias vi reducerer.

Svaret på udfordringen er derfor at have fokus på alle led i en dataindsamling. Sikre repræsentativiteten i udvælgelsen, og derefter reparere og standardisere det, der er muligt. Hele tiden afvejet i forhold til ressourcer. Så man skal have fokus på alle dele af dataindsamlingen og sætte særligt ind der, hvor det svageste led er, da ingen dataindsamling er bedre end sit svageste led.

Kvalitet af en dataindsamling

Hvor retvisende en stikprøve er er afhængig af en lang række faktorer. De tre mest kritiske faktorer er normalt: 1) kvaliteten af udtrækspopulationen og stikprøveudvælgelsen, 2) hvor gennemtestet og brugervenligt spørgeskemaet er, samt 3) bortfaldet i dataindsamlingen.

I denne artikel er der særligt fokus på reparationen af bortfaldet – punkt 3 ovenfor. Det betyder ikke at forebyggelse ikke er det vigtigste, men at begge dele er nødvendigt. Som udgangspunkt antages det, at der er tale om en stikprøve med universelt repræsentativitet, hvis alle svarer. **Universel repræsentativ** betyder, at stikprøven er trukket med en kendt positiv sandsynlighed i hele populationen, og de efterfølgende svar i estimationen er vægtet med det inverse af denne sandsynlighed – det vil sige med hvor mange personer 'svaret' repræsenterer i hele populationen. Kun en sådan *probability samling* i hele populationen og efterfølgende vægtning for udvalgssandsynligheden sikrer, at alle estimater lige så godt kan være over som under det rigtige, at de i gennemsnit er det, og sikkerheden bliver herfor større, hvor flere der udvælges. Dvs. medfører at forudsætning (2) er opfyldt for alle hypoteser i undersøgelsen.

En udvælgelse med forskellige udvalgssandsynligheder opstår let i praksis. Hvis der fx udvælges tilfældige telefonnumre, afhænger udvalgssandsynligheden af, hvor mange telefoner personen kan træffes på. Og hvis det er en to trins-udvælgelse af husstande og derefter udvælgelse af én person i husstanden af hvor mange personer, der er i husstanden. I praksis er det i en generel befolkningsundersøgelse kun muligt at sikre ens udvalgssandsynlighed for alle, hvis der vælges simpelt tilfældigt i CPR registret, som alle private og offentlige virksomheder har adgang til at bruge til statistiske undersøgelser. Hvis der er forskellig udvalgssandsynlighed, skal der vægtes herfor – det kaldes i litteraturen *designvægten* eller den *direkte vægt*. Hvis der ikke er en sådan vægt tilknyttet svarene i en stikprøve, hvor ikke alle har samme udvalgssandsynlighed, er der ikke tale om en repræsentativ undersøgelse, og forudsætning (2) i den simle test er ikke opfyldt.

Reparation af bortfald

Udover en designvægt kan det være nødvendigt med en bortfaldsvægt, hvis bortfaldet ikke er ens på alle faktorer, for at understøtte at mulig bias reduceres, og forudsætning (2) er opfyldt så langt, som det er muligt. I praksis vil bortfaldet altid være skævt, og generelt er det fx de mest integrerede, informationssøgende, erhvervsaktive, uddannede og ældre personer, der svarer mest. Det kalder på en vægt, der standardiserer for disse forhold, med mindre man ønsker en for positiv beskrivelse af mange forhold i populationen.

De unikke registre i Danmark giver internationalt en enestående mulighed for at håndtere skævheden pga. bortfaldet, der standardiserer stikprøven i forhold populationen mht. signifikant registervariabler. Oprindeligt har man tidligere standardiseret med vægte for alder, køn og geografi, fordi det var det eneste, man kendte for alle. Denne vægtning var langt fra tilstrækkelig, men det bedste man tidligere kunne præstere. I dag har vi langt bedre muligheder: indkomst, sociale forhold, nationalitet og uddannelse findes i registrene og er fx normalt langt mere signifikante, og derfor bedre faktorer end køn, alder og region til at forklare bortfaldet.

En vægt efter dataindsamlingen er gennemført kaldes for *efterstratifikation*. Som navnet antyder opdeler man populationen i en række undergrupper (strata) eller tabeller efter dataindsamlingen. Og i hvert stratum er vægten 'antal svar divideret op i populationen'.

Traditionel vægtning af middelværdier

I en simpel tilfældig stikprøve kan man ønske at vægte (standardisere) efter uddannelse. Man tildeler derfor en vægt efter uddannelse, fx:

Høj: $$W_1 = \frac{N_1}{n_1} \qquad (3)$$

Lav: $$W_2 = \frac{N_2}{n_2}$$

Hvor N_1 hhv. N_2 er antal med høj hhv. lav uddannelse i populationen og n_1 og n_2, med tilsvarende notation, den opnåede stikprøve (antal svar) for hhv. lavt og højt uddannede.[1]

Vægtene i formel (3) genskaber de oplysninger i populationen, man har stratificeret efter. Hvis man fx summer vægtene for de n_1 højtuddannede i stikprøven giver det præcist N_1, der er det samlede antal højtuddannede i populationen.

[1] Vægtene i formel (3) svarer til designvægtene, hvis n_1 og n_2 er de oprindelige stikprøver før bortfald, og de kan være forskellige, hvis man har valgt flere relativt i den ene gruppe frem for i den anden gruppe.

Med disse vægte kan man estimere fx den samlede lønsum i hele populationen, hvor variablen y måler lønnen for de n personer i stikprøven:

$$\hat{Y} = \sum_{j=1}^{2} \sum_{i=1}^{n_j} W_j y_{ji} \qquad (4)$$

Denne opregning kaldes også ***model-assisted*** frem for *model-depended* i international litteratur, da vægtningen alene standardiserer for skævheder i stikprøven, og ikke ændre svarene. Hvis man fx skulle estimere den gennemsnitlige løn, og bortfaldet er størst for lavt uddannede, ville det ikke være retvisende, at lade skævheden i stikprøven påvirke estimatet af den gennemsnitlige løn.

Model-assisted kontra model-depended

Begrebet *model-assisted* om vægtning for bortfald skyldes en grundlæggende egenskab med vægtningen. Antag man skal estimere den gennemsnitlige løn for hele befolkningen, og der er skævt bortfald, og flere højtuddannede har svaret, ved vi at estimatet for løn har en skævhed og er for højt, fordi der er flere højt uddannede i forhold til lavt uddannede, der har svaret. Dette kan vi standardisere på plads med vægtene. Vi kan selvfølgelig ikke løse andre skævheder i dataindsamlingen end denne fejlkilde med vægten. Men hvorfor skulle det afholde os for at rette noget på plads, som skal være på plads, hvis den statistiske analyse og estimatet for den gennemsnitlige løn skal give mening. Hvis man ikke vægter, er man nødt til at antage, at uddannelse ikke betyder noget for løn og det skæve bortfald derfor heller ikke betyder noget.

Der er andre måder at reparere for bortfald på, der er model-**depended** og ikke kun model-assisted. Fx imputering af de manglende svar. Her skal man også vælge faktorer at imputere efter (modellen), og hvis man fx imputerer efter uddannelse, kan de manglende svar for fx højt uddannede imputeres ved gennemsnittet for de, der har svaret. Det vil faktisk give samme estimat som i formel (4), hvis selve stikprøven før bortfald er repræsentativ. Men estimatet af variansen vil blive for lille, da de manglende svar har fået samme værdi, som gennemsnittet variansen beregnes omkring, og afhænger derfor af (depended) af imputeringen. Og hvis bortfaldet også er skævt mht. køn (flere kvinder har svaret), og vi ikke har imputeret efter denne faktor, vil lønforskellen mellem mænd og kvinde fremtræde mindre, dvs. med en bias, fordi kvinder og mænd, der ikke har svaret, har fået imputeret sammen løn. Det vil ikke ske, hvis vi kun har vægtet efter uddannelse og ikke køn - og sammenligner køn. For når vi deler op efter køn, estimerer vi jo for hvert køn for sig (og så er det ligegyldigt om der er vægtet for køn eller ej). Så imputering kan vaske forskelle bort – vægten bibeholder dem, men standardiserer dem ikke generelt. Derfor er vægtning kun model-assisted. Imputering kan give model bias, hvis ikke alle signifikante faktorer er inddraget i imputeringen. Vægtningen ville blive bedre, hvis der var blevet vægtet af

alle vigtige signifikante faktorer, men estimaterne får ikke ny bias, hvis man ikke gør det og ikke har standardiseret efter alt.

Estimation af variansen

Inden vi kigger på variansen ved vægtning, så vi kan få formel (1) med den basale test helt på plads, skal vi have lidt flere definitioner. Vægtene i formel (4) summer til populationen. I statistiske analyser ønsker man ofte, at vægtene summer til den faktiske stikprøve, så variansberegninger afspejler den opnåede stikprøvestørrelse. Dette sikres let ved at korrigere vægtene med det samme, nemlig n/N:

Høj: $$g_1 = \frac{N_1}{n_1} \bullet \frac{n}{N} \quad (5)$$

Lav: $$g_2 = \frac{N_2}{n_2} \bullet \frac{n}{N}$$

$$\sum_{i=1}^{n_1} g_1 + \sum_{i=1}^{n_2} g_2 = n \quad (6)$$

Hvor n er den opnåede stikprøve og N populationen. G-vægtene har gennemsnit 1.

Disse g-vægte kan let bruges til at estimere den gennemsnitlige løn da:

$$\hat{y} = \hat{Y} / N = \sum_{j=1}^{2} \sum_{i=1}^{n_j} g_j y_{ji} / n \quad (7)$$

Alle programpakker kan i dag gennemføre en statistisk analyse og lave tabeller med vægtede data. Vægte for at håndtere skævheden pga. bortfaldet påvirker også variansen – dvs. nævnere i formel (1). Som udgangspunkt betyder vægtene, at variansen vil kunne øges, hvilket er den pris, man må betale for at håndtere skævheden pga. bortfaldet. Men vægtning med registervariabler betyder også, at stikprøven bliver mindre usikker, svarende til *granden af forklaret variation* (R^2), som opregningsvariablerne. Denne gevinst kan være ganske pæn.

Gennemsnittet af g-vægtene i stikprøven har gennemsnit 1. Når der efterstratificeres er variansen for fx den gennemsnitlige løn i Danmark:

$$V(\bar{y}) = \left(\frac{N_1}{N}\right)^2 \frac{\sum_{i=1}^{n_1} (y_{1i} - \bar{y}_1)^2}{n_1 - 1} \cdot \frac{1}{n_1} + \left(\frac{N_2}{N}\right)^2 \frac{\sum_{i=1}^{n_2} (y_{2i} - \bar{y}_2)^2}{n_2 - 1} \cdot \frac{1}{n_2} \quad (8)$$

Denne variansformel er den sædvanlige for stratificerede stikprøver. Formelen kan med stor tilnærmelse (svarende til n-1/n=1) omformuleres til:

$$V(\bar{y}) = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j} g_j^2 (y_{ji} - \bar{y}_j)^2}{(n-1)n} \qquad (9)$$

Bemærk at $\sum_{j=1}^{2}\sum_{i=1}^{n_j} g_j = n$ (formel 6) så man kunne have formuleret formel (9) som:

$$V(\bar{y}) = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j} g_j^2 (y_{ji} - \hat{y})^2}{(\sum_{j=1}^{2}\sum_{i=1}^{n_j} g_j - 1)n} \qquad (9a)$$

Formel 9 og 9a har to konsekvenser. Fordi g-erne sættes i anden potens vil dette første produktled i tælleren normalt give en variansforøgelse.[2] Men da efterstratificeringen har fjernet variation mellem høj og lavt uddannedes løn, da variationen beregnes indenfor høj hhv. lav uddannede, vil det andet led i tælleren omvendt give en variansreduktion. Denne reduktion svarer til, hvor meget uddannelse forklarer mht. variationen af løn. Ellers sagt på en anden måde – variansforbedringen er $1-R^2$. At man skal sætte vægten i anden potens skyldes, at variansen er den gennemsnitlige kvadratafstand: $V(g \cdot X) = g^2 \cdot V(X)$.

Formel (9a) svarer til den måde, man beregner variansen, hvis man fx i proc surveymeans i SAS angiver weigt=g. Her er formelen nemlig:

$$V(\bar{y}) = \frac{\sum_{j=1}^{2}\sum_{i=1}^{n_j} g_j^2 (y_{ji} - \hat{y})^2}{(\sum_{j=1}^{2}\sum_{i=1}^{n_j} g_j - 1)n} \qquad (10)$$

Når man sammenligner formel (9a) med formel (10) fremgår det, at man i formel (10) ikke får gevinsten svarende til graden af forklaret variation mht. registervariablerne, fordi variansen ikke beregnes omkring de to lokale estimater, men alene betaler prisen for de varierende g-vægte i anden potens. Det er det der sker, hvis man i proc surveylogistics eller tilsvarende procedurer i SAS angiver vægten g, uden at angive strata. Parametrene vil være rigtige, da g-vægtene kontrollerer herfor, men variansen er normal lidt for stor. Det betyder populært sagt, at man tester konservativt – der skal mere til for, at forskelle bliver signifikante. Men når de er signifikante, kan man være sikker på de også er signifikante. Hvis man vil være helt skarp, kan man i proc surveyfreq, surveymeans eller surveyreg bruge option **strata** og derved – udover at kontrollere for skævheden ved bortfaldet – også at opnå den mulig variansgevinst i formal 91. Det kan klart anbefales 😊

[2] Hvis n=2 og g-erne er 1,5 hhv. 0,5: $1{,}5^2+0{,}5^2 = 2{,}5$ og $1^2+1^2 = 2$. Så vægtene øger her variansen med 25%.

Det er derudover vigtigt, at man fx bruger survey procedurerne i SAS eller svy i STATA, og hermed regner rigtigt mht. både middelværdi og den grundlæggende egenskab ved vægte i beregningen. Hvis man ikke gør det sidste sætter man ikke gerne i anden potens i formlerne 9 og 10, dvs. man har en for lille varians i sin test. At teste med en for lille varians fører til konklusioner, der reelt ikke er signifikante. Kan klart ikke anbefales.

Efterskrift

Ovenående tilgang til vægtning kan udvides til mange faktorer og kontinuerte variabler og komplekse design. Det er belyst i de kilder, der står nedenfor.

Kilder

Andersson, C. (2004). *CLAN97 v3.1 Supplement to 'A usersguide to CLAN97'.* Statistics Sweden.

Andersson, C. and Nordberg, L. (1998). A User's Guide to CLAN97. Statistics Sweden.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling, Springer-Verlag.

Stine Fangel, Peter Christian Linde og Brian Larsen Thorsted (2007). *Nye problemer med repræsentativitet i surveys, som opregning med registre kan reducere*, Metode og Data nr. 93 - 2007, side 14-26.

Market shares of beer, wine and spirits in Denmark

Anders Milhøj

Department of Economics University of Copenhagen

anders.milhoj@econ.ku.dk

Summary

In this paper data on the Danish consumption of beer, wine and spirits is studied in order to identify changes in the market shares of each of these types of alcohol caused by trends in taste and changes in the price relations among the alcohol.

Plots of the data immediately show that the market shares have changed dramatically. Part of these changes are due to the rather large changes in the relative prices mainly due to the major tax reductions for beer and wine in 1991 and 1992 and a tax reduction in 2003. But also changes in taste and drinking habits might have caused the development, as well as influence by fluctuations in the border trade with Germany, that were the main reasons for the tax reductions.

Moreover, it is now agreed (Finanslovsaftalen 2019) to reduce the taxation on beer and wine each year in the period 2019 - 2022, so the expected impact of these price changes is studied in the last section of this paper.

The Danish alcohol market

The Danish consumer taxation on alcohol has changed frequently. In the eighties the taxation was adjusted, mainly upwards, many times, and in the nineties the taxation on beer and wine, but not on spirits, was reduced twice, July 1.1991 and October 1. 1992. These reductions were quite drastic: approximately 10% in 1991 and as much as 25% for some low price brands of beer in 1992. By October 1. 2003 the taxation on spirits were similarly reduced drastically. These drastic changing prices give a unique opportunity to study the effects on drinking habits caused by price changes.

The reason for the ongoing process of reducing taxes is the high level of German border trade. It is a commonly accepted fact, that literally no one in the southern parts of Jutland buy wine and beer in Denmark and also Danes living far away from the German border go on regular trips to Germany mostly to buy alcohol. Long ago the Danish government tried to control this traffic with various kinds of restrictions on the imported quantities, but as such restrictions are now inconsistent with the intentions of

the European Community, they had to be relaxed. As alcohol is considered as the main reason for border trading, the retailers especially in southern Jutland often urge for a solution.

The tax reductions are of course valid for the country as a whole including the island of Zealand with the Copenhagen area, where border trade of any significant magnitude of geographical reasons is more expensive and time consuming.

On the other hand, the border trade with Sweden and to some extend also Norway goes the other way. Especially after Sweden joined the European Union, which weakened the trade restrictions, the export of alcohol by travellers from Denmark to Sweden is considerable. Some of the fluctuations in the total sales of alcohol could be explained by changes in the border trade. Especially the raise in the total quantity sold in the period 1991 to 1995 could be caused by reduction in the border trade with Germany combined with a raise in the border trade with Sweden caused by the huge tax reductions i Denmark on beer and wine in 1991 and 1992.

The total quantity of alcohol sold in Denmark has been remarkably constant for the years 1981 - 2017 considered in the present paper. Of course the population size has increased, the age distribution has changed towards more elderly people and the Muslim (non drinking) population minority has raised sharply. But as the analyses are based on market shares and not actual sales numbers it is unnecessary and to some extent even misleading to normalize by population size.

The present paper could be seen as an updated version of some previous papers, for instance Milhøj (1993 and 1995 in Danish and laCour & Milhøj(2011) in order to study how the market has reacted in the years thereafter - especially the market share of spirits changed after the tax reduction on October 2003.

Data

The Danish Statistical Bureau provides yearly data on the sales of alcohol measured in 100% alcohol. In this paper we denote these variables

BSALE Volume of beer sales in 1000 litres 100% alcohol

WSALE Volume of wine sales in 1000 litres 100% alcohol

SSALE Volume of spirits sales in 1000 litres 100% alcohol.

The total sales of alcohol measured in 100% alcohol is then defined as

TOTSALE = BSALE + WSALE + SSALE.

As a part of the monthly reported Danish Consumer Price Index the Statistical Bureau gives prices for the three types of alcohol:

BPRICE Index numbers for the price of beer

WPRICE Index numbers for the price of wineSPRICE Index numbers for the price of spirits.

The consumer price indices, which are based on 1980 = 100, give the prices as they are met by the consumers, with taxes included.

Both the sales and price series are aggregated, such that eg. beers of various brands, strengths and prices are considered as one. Moreover, the consumer price index is based upon "normal" prices from a sample of Danish retail shops, which do not take temporary price reductions etc. into account. It is however clearly possible, that the price driven substitution in consumption among the three types of alcohol is more pronounced for low price brands of beer, wine and spirits, than for more expensive brands.

The main subject of the present analysis is to study the price driven substitution, and for this reason the market shares of the three types of alcohol are considered as the variables of interest. Here we denote these variables:

LBTOT = log(BSALE/TOTSALE)

LWTOT = log(WSALE/TOTSALE)

LSTOT = log(SSALE/TOTSALE).

These variables are the logarithmically transformed market shares of the three types of alcohol in the total Danish alcohol market. In the models these variables are explained by a linear regression model based on the log-transformed relative prices

LOGPBW = log(BPRICE/WPRICE)

LOGPBS = log(BPRICE/SPRICE)

LOGPWS = log(WPRICE/SPRICE).

A graphical analysis

In Figure 1 the total sale of alcohol (measured in litres of100% pure alcohol) in Denmark in the years 1978 to 2017 is shown. As can be seen from this diagram the sale was increasing in the period 1978 - 1983, somewhat declining in the late eighties,

again increasing to a top in 1996 and since then again somewhat declining. The level seems rather constant considering the whole period 1978 - 2017 at about 45 - 50 mill. litres of 100% pure alcohol. In fact, sales seem to decline - perhaps unexpected as the Danish population has increased by more than 10% in this period. igure 1 also shows the sales of the three types of alcohol in the years 1978 - 2017. Most striking is that the sales of wine have increased steadily over the whole period showing more than a doubling. The sales of beer were almost constant until 1996 and have been steadily declining since then. The sales of spirits have fluctuated until the beginning of the eighties and then dropped, most dramatically in 1993, then stabilised at a low level for many years, but since 2003 sales of spirits increased again.

Figure 2 shows yearly series of the market shares of the three types of alcohol, that is the variables LBTOT, LWTOT and LSTOT before the log-transformation. The market shares of spirits as a part of the total Danish alcohol sales has decreased from 16% in the eighties, then decreased to 10%, but then increased to again 16% in recent years. The market shares of beer show some fluctuations in the eighties and nineties, but in recent years the shares have been steadily declining from 65% to a level around 38% for the previous 8 years. The market shares of wine on the other hand has been steadily increasing, in fact it has more than doubled from 19% to 46% of the total Danish consumption of alcoholics. The fluctuations in 2010-2011 could be an effect on changes in the taxation on beer, but to some extend it due a technical regulation of the calculation of the alcohol contents in wine.

Figure 3 presents yearly price series for the three types of alcohol indexed by 1980 = 100. Of course, inflation gives steady upward trends in all price series, but moreover effects of the frequent minor adjustments of the taxation on alcohol in the years 1982 - 1986 imply a larger volatility in the price series. The tax reductions for beer and wine are clearly seen.

A similar picture is seen in Figure 4, which presents the price for the three types of alcohol to the general consumer price index. In these graphs all effects of inflation and general price changes are removed by the normalization. The price relation between beer and wine has been almost constant until the mid-eighties, but since then beer has become relatively cheaper than wine. The price reductions of beer and wine relative to spirits caused by the tax reductions in 1991 and 1992 are clearly seen. The price of spirits has increased considerably compared to the prices of both beer and wine, but due to the tax reduction in 2003 spirits became relatively cheaper. As the series presented all have the value 1980 = 1 as the basis year, it is seen that the prices of all types of alcohol have decreased dramatically.

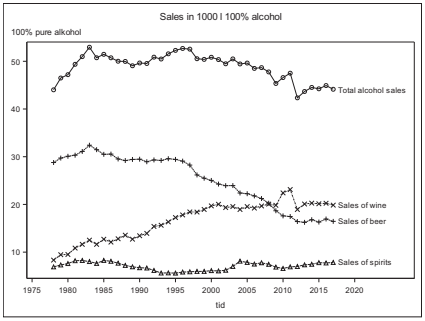


Figure 1 Sales in 1000 l 100% alcohol

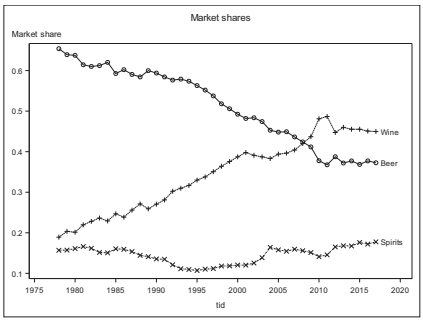


Figure 2 Market shares

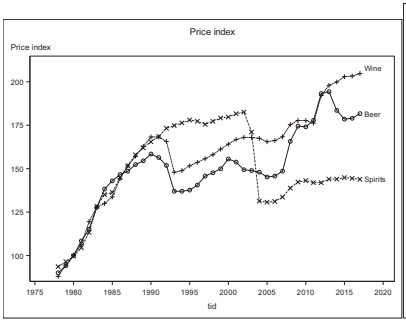


Figure 3 Price indeces

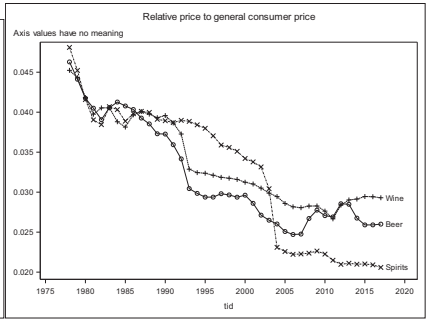


Figure 4 Relative general consumer price

The statistical model

In this section the basic statistical model is presented. The model is constructed in a way that is particularly well suited for the study of the effects on the market shares of wine, beer and spirits in the total alcohol market caused by changes in their relative prices. It has been demonstrated by the graphical analyses that the quantity of alcohol sold in Denmark has been remarkably constant in the whole period 1981 to 2002. Although the total alcohol sales are constant, the quantity sold of each of the three types of alcohol has changed dramatically giving a clear trend towards wine at the expense of beer and spirits. It seems reasonable to suggest for instance that the Danes have developed a stronger taste for wine partly because of changes in eating habits and partly because of increased travelling in southern Europe. On the other hand, the changes in the price relations could also be the reason for the changes in market shares.

The question of whether these phenomena could be explained by changes in the price relations or should be seen mainly as a trend in habits is the focus of the statistical models in the present paper. To this end linear regression models are used with the market shares LBTOT etc. as response variables and the relative prices LOGPBS etc. as explanatory variables. In the statistical models the development in the level of the market shares is described by a trend. As seen from the plots in figure 4 this trend is possibly not linear for all price ratios and also quadratic trend curves are applied. To this end two explanatory variables, TREND, which simply equals the observation number and QTREND, equal to the observation number squared, are included in the models.

With relative prices towards both beer and wine and the trend as the explanatory variables, the basic model for the market shares of spirits is written as

$$LSTOT = \beta_{INT} + \beta_{BS}LOGPBS + \beta_{WS}LOGPWS + \beta_{TREND}TREND + \beta_{QTREND}QTREND.$$

In this model the market shares of spirits (and not the actual sales volume of spirits) is used as the response variable. Therefore, changes in the total sales of alcohol have no effect on the response variable as long as the sales for the three types of alcohol change proportionally. Many factors that could possibly influence the total sales of alcohol such as income, general taxation, prices of other goods such as housing prices, will under the condition of proportionality have no influence on the market shares. Also changes in the population size could under these circumstances be excluded from the model. If the effects of external factors on the three types of alcohol are however not proportional, the effects are probably included in the trend terms. The important point is, that the trend terms include the shift in drinking habits from year to year caused by the fact that the older diminishing generations prefer beer or aquavit, while the new growing younger generations prefer wine.

Fitted models

In this section the results of estimations of parameters from the various versions of the basic model are presented, as fitted models using the estimated parameters. Moreover, some technical comments are given, while the concluding section is devoted to a more intuitive discussion of the results.

In all of the models presented, the actual numerical value of the intercept is of no direct interest and hence is omitted. Moreover, as some of the estimated parameters are insignificant, significant values using a two sided test on level 5% are identified by bold faces numbers.

The fitted relations are of course checked by the usual control procedures for regression models. Problems with autocorrelation in the models are explicitly dealt

with in the presentation of the models below. All models present, as is usual for time series data, a good fit with large values of R^2.

Fitted models for the market shares of beer

The estimated basic model for the market shares of beer is

LBTOT = INT **- 0.29** LOGPBW **- 0.20** LOGPBS **- 0.015** TREND.

In this relation some autocorrelation is present with a first order autocorrelation equal $\rho_1 = 0.44$. If the relations are re-estimated using maximum likelihood techniques allowing for a first order autoregressive remainder we get

LBTOT = INT - 0.18 LOGPBW - **0.18** LOGPBS - **0.0.15** TREND, $\varphi_1 =$ **0.54**

where φ_1 is the autoregressive parameter. Note that the price ratio beer/wine now losses its significance.

In summary, the analysis demonstrates that the development in the market shares of beer includes a downward trend. Moreover, both price ratio are significant.

Fitted models for the market shares of wine

When the basic model is fitted to the market shares of wine, only the trend variable is significant, while the relative beer/wine price as well as the relative wine/spirits price give no significant contribution to the model,

LWTOT = INT - 0.23 LOGPBW **- 0.42** LOGPWS **+ 0.028** TREND.

In this model no severe autocorrelation is present, $\rho_1 = 0.18$ and the inclusion of a quadratic term in the trend offers no improvement. Note that the price ratio beer/wine i sin significant in accordance to the similar result above.

The fitted model using only the price ratio of wine to spirits and the trend is

LWTOT = INT **- 0.39** LOGPWS **+ 0.027** TREND.

Again the clearly significant trend dominates, while the price ratio of wine to spirits is somewhat significant. Hence it must be concluded, that the market shares of wine mainly depend on a linear trend, but that the upward pointing market shares of wine also to some extent is due to the changes in the price ratio leading to relatively cheaper wine.

Fitted models for the market shares of spirits

For the market shares of spirits the trend is of importance, but the price ratios also enter the model. Using both price ratios and a linear trend variable we find

$$\text{LSTOT} = \text{INT} + 0.17\ \text{LOGPBS} + \mathbf{0.99}\ \text{LOGPWS} - \mathbf{0.009}\ \text{TREND}.$$

This relation presents some autocorrelation, $\rho_1 = 0.63$. When estimated allowing for a first order autoregressive remainder term, the model becomes

$$\text{LSTOT} = \text{INT} + 0.25\ \text{LOGPBS} + \mathbf{0.63}\ \text{LOGPWS} - 0.006\ \text{TREND},\ \varphi_1 = \mathbf{0.83}$$

The conclusion is that the price ratio of wine to spirits enters the relation with a highly significant coefficient.

Though it seems from these relations that only the beer/spirits price ratio and not the wine/spirits price ratio is of importance, this conclusion could be a statistical artefact due to multicollinearity. But as the two price ratios have been parallel in the years under study, it is hard to separate the effects of changes in the beer/spirits price ratio from the effects of changes in the wine/spirits price ratio.

Guesses for markets shares the next three years

At the moment of writing, December 2017, yearly data for all of 2018 is not yet at hand, but no changes in taxation occurred and monthly data for sales point at no remarking changes. For the next four years however taxation changes are planned for wine and beer, but not for spirits, in order to reduce border trade.

Type	Total revenue 1000 kr.	Reduction 2019	Reduction 2020	Reduction 2021	Reduction 2022
Beer	910,689	70,000	90,000	90,000	85,000
Wine	1,720,873	25,000	35,000	30,000	30,000
Spirits	1,206,424	-	-	-	-

For beer this corresponds to a reduction of 0.40 kr as the end result in 2022 for an ordinary 0.33 cl bottle or can 4.6% beer. But order to calculate how this tax reduction affects the consumer price index many assumptions are needed; for instance that the price of beer of course varies as to quality and place of consumption which could be at an expensive restaurant or a cheap supermarket. For wine it is even harder because the

taxation of wine is not exactly depending on the alcohol content, but instead on different interval corresponding to various categories of wine. But for a bottle 0.75 l common red or white wine the total reduction corresponds to 0.76 kr as the end result in 2022.

So it is of course impossible to end up with a definite result on future values of the relative prices used here, but an attempt is presented in Figure 5 which is a continuation of Figure 3. Here also the prices for 2023 to 2030 in case of no further taxation changes come up after 2022.

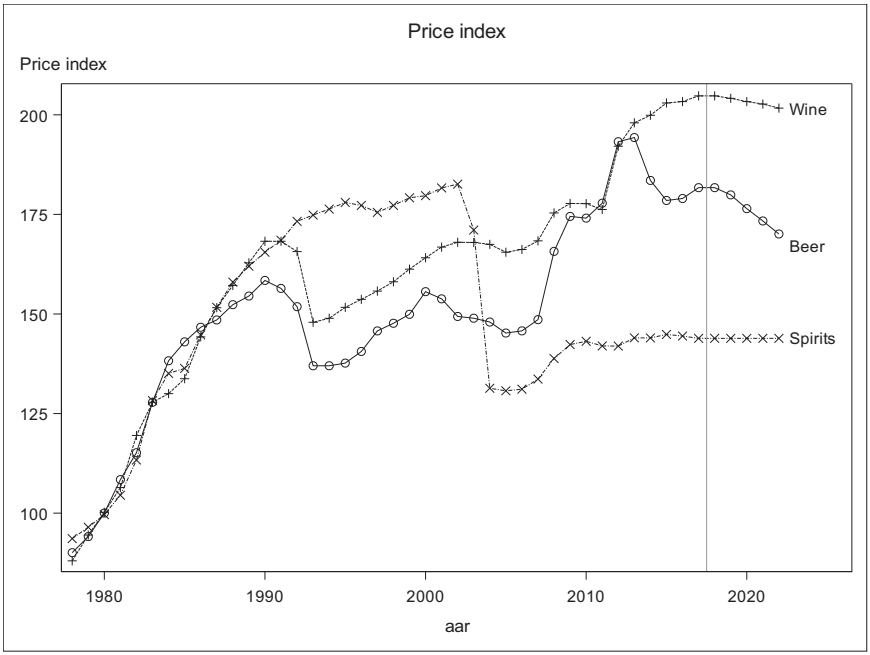


Figure 5 Expected future prices.

Figure 6 and 7 then present forecast of future market shares according to the models presented in the present paper. The results reflect to some extend that the proposed the taxation changes favours both wine and beer to spirits and furthermore favours beer more than wine. The forecasts in Figure 6 reflects that the trend is dominating; especially for the market shares of wine, while Figure 7 gives the forecasts when the trend is excluded from the model.

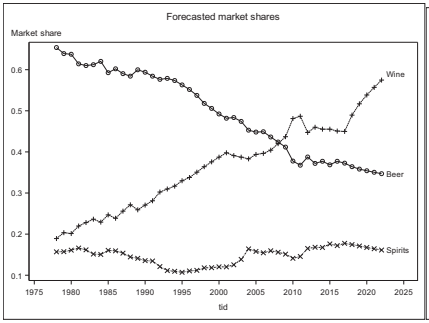


Figure 6 forecasts with a trend

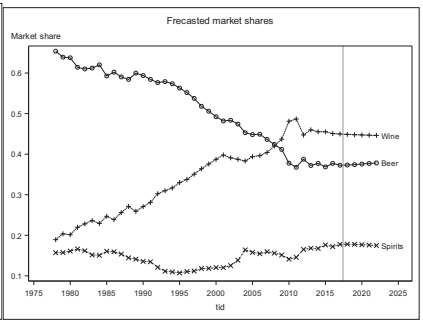


Figure 7 Forecasts without a trend

Conclusion

In all relations the trend gives a significant contribution to the model fit. For the sales of wine, the linear trend parameter $\beta_{TREND} = 0.028$, when transformed back using the exponential function nearly corresponds to a 3% increase in the market shares of wine.

For the market shares of spirits the coefficient to the wine/spirits price ratio is around 0.63. This means that a 10% reduction in the relative beer/spirits price approximately causes a reduction of 6.3% in the market shares (measured in percent points) of spirits. The similar interpretation of the relation for the market shares of wine uses the estimated value of β near 0.39. This means that a 10% reduction in the relative beer/spirits price causes a reduction by approximately 3.9% in the market shares of wine. For beer the results are weaker withe confidents around 0.2 saying that the relative prices do not matter that much compared to downward trend.

References

la Cour, L. F and Milhøj, A. (2011): The Sale of Alcohol in Denmark - Recent Developments and Dependencies on Prices/Taxes'. Applied Economics 2009 vol 41 (9), 1089-1103.

Milhøj, A.: Virkningen af afgiftsnedsættelser på øl- og vinsalget i Danmark. Nordisk Alkoholtidsskrift 10 (1993) 319 - 329.

Milhøj, A.: Udviklingen i danske familiers øl- og vinindkøb 1990 - 1993. Nordisk Alkoholtidsskrift 12 (1995) 181 - 193.

Budgetloven: strukturel budgetbalance, finanspolitisk holdbarhed og asymmetrisk udgiftsloft

Jesper Jespersen, Roskilde Universitet. e-mail: jesperj@ruc.dk

'Budgetloven blev vedtaget af et bredt flertal i Folketinget i maj 2012 og har i praksis haft virkning fra finansåret 2014.

Budgetloven har tre/fire overordnede stopklodser med hensyn til styring og begrænsning af de offentlige udgifter.

Med Budgetloven blev begreberne 'strukturel budgetbalance', 'finansiel holdbarhed' og 'udgiftslofter' en del af den politiske diskurs/juridiske virkelighed.

Problemerne hermed er mange. Det største er formentlig, om de tre begreber er relevante, hvis formålet med finanspolitikken er at indfri divergerende politiske præferencer og sikre en stabil makroøkonomiske udvikling på hhv. kort og længere sigt? Det forudsætter en hidtil uset politisk enighed og hertil en hidtil uset stabilitet i den private reale og finansielle markedsøkonomi. Forhold der – mildt sagt - sjældent genfindes uden for generelle ligevægtsmodeller.

Hertil kommer de statistiske problemer, der knytter sig til opgørelsen af de 'nye' begreber, idet

1. de ikke er entydigt bestemt, da definitionerne varierer,

2. rammen for næste års finanspolitik er baseret på hypotetiske, modelberegninger, der rækker 5-7 år (strukturelt budget) og 50 år (finanspolitisk holdbarhed) ud i fremtiden!

3. ydermere er regelsættet knyttet til disse – mildt sagt usikre beregninger - asymmetrisk i sin udformning og dermed i dets makroøkonomiske effekt: straf følger hvis loftet overskrides, ingen belønning (det vil sige ingen mulighed for at anvende et underforbrug det følgende år) hvis loftet 'underopfyldes' – uanset at der er en betydelig kassebeholdning.

– giver det mening?

1. Finanspolitisk holdbarhed.

Finanspolitikken skal ifølge udgiftsloven være langtidsholdbar. 'Det betyder groft sagt, at vi skal kunne opretholde velfærdsstaten, som vi kender den i dag - uden at det er nødvendigt at hæve skatterne', DØR, 2015

Beregningen af den finanspolitiske holdbarhed er uhyre følsom over for de til grundlagte forudsætninger vedr. fertilitet, indvandring, erhvervsfrekvens, pensionsalder, realrente, produktivitet i både privat og offentlig sektor, samt den valgt tidshorisont, ofte benyttes en 50 års periode, hvor modelberegningernes usikkerhed vil spille en dominerende rolle. Hvis nogen skulle være i tvivl herop, kan jeg anbefale at læse de *perspektivplaner*, der

blev udarbejdet i begyndelsen af 1970erne, der få år senere – i lyset af bl.a. energi- og beskæftigelseskriserne – viste sig ubrugelige til at løse de presserende samfundsøkonomiske problemer.

Som det vil blive uddybet i de efterfølgende afsnit er begrebet finanspolitisk holdbarhed på grænsen af det meningsløse på grund af den usikkerhed, der knytter sig til beregningen både hvad angår de fremtidige samfundsøkonomiske udfordringer, de politiske prioriteringer og den anvendte makroøkonomiske model. Tilsammen gør disse tre faktorer beregningen mere vild- end vejledende.

'At finanspolitikken er holdbar vil overordnet sige, at den politik, der planlægges inden for den mellemfristede planlægningshorisont, kan videreføres uændret derefter, uden at det fører til, at den offentlige gæld vokser ukontrolleret på langt sigt'.
https://www.fm.dk/oekonomi-og-tal/oekonomisk-analyse/2017/holdbarhed-og-generationsmaessig-neutralitet
Finansministeriet fortolker generationsneutralitet, som uændret skattetryk i fremtiden. Dette er overraskende, idet Finansministeriet i andre fremskrivninger har beregnet at BNP pr. capita om 40-50 år vil være det dobbelte af det aktuelle, se f.eks. Finansministeriet, xx. Det er om noget *et politisk statement*, at fremtidige generationer (for ikke at nævne nutidige) i takt med velstandsstigningen *ikke* skulle ønske at allokere en større del af velstands***stigningen*** til løsning af kollektive udfordringer, som for eksempel klimaudfordringen og den stigende (relative) fattigdom.

2. *Strukturel offentlig saldo*: Et mål for finanspolitikkens retning, styrke og hensigtsmæssighed?

'Budgetloven siger, at det strukturelle underskud højst må udgøre ½ pct. af BNP. Den strukturelle saldo er et mål for, hvordan den offentlige saldo ville se ud, hvis situationen på de finansielle markeder, konjunktursituationen mv. var normale'. DØR, 2015

Med budgetloven er finanspagtens krav til balance på de offentlige finanser implementeret i dansk lovgivning. Det årlige strukturelle offentlige underskud må således højst udgøre ½ pct. af BNP'. (Finansministeriets hjemmeside).

Da finanspagten er en mellemstatslig aftale, der ikke er en del af EU-traktaten, og som f.eks. Tjekkiet og Storbritannien har holdt sig uden for, kan det undre at den danske, dengang socialdemokratisk ledede, regering underskrev finanspagten uden en grundig politisk og fagøkonomisk diskussion.

En sådan diskussion ville have afsløret, at den fagøkonomske vurdering af det finanspolitiske råderum er stærkt delte, se Jespersen (2018). Det er spørgsmålet om den private sektors 'self-adjusting' egenskaber, der står i centrum for denne diskussion. Yderligt suppleret med 'the Treasury View on Public Finance', at finanspolitikkens primære mål er at sikre ligevægt på det offentlige budget, der er indeholdt i stabilitetspagten (1996) og som blev yderligere opstrammet i Finanspagten (2012)[1].

[1] Retfærdighedsvis bør det også nævnes, at der selv I EU-systemet er delte meninger om det hensigtsmæssige i at have 'ligevægt på de offentlige budget' som primært finanspolitisk mål, jfr. Romani Prodi, EU-kommissionens

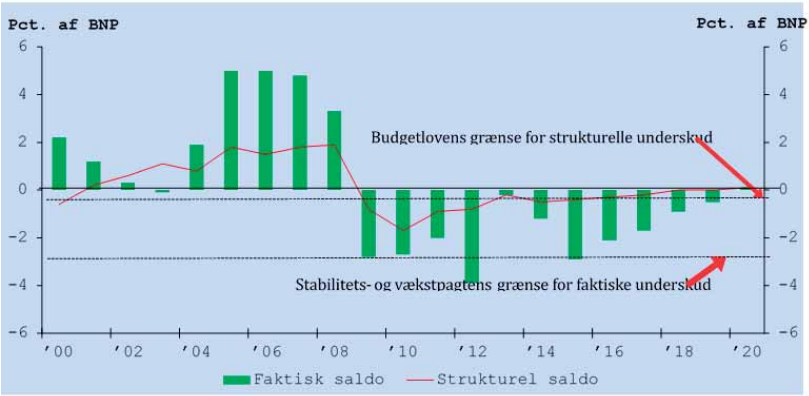
I praksis er den strukturelle saldo ligesom finansiel holdbarhed baseret på beregninger gennemført ved brug af Finansministeriets makroøkonomiske model med den usikkerhed (teoretisk (generel ligevægt) og empirisk (delvis kalibreret), som det i sig selv indebærer. En hypotetisk beregning gennemføres, hvori der manipuleres med efterspørgslen i den private sektor, indtil produktion og beskæftigelse svarer til hhv. 'fuld kapacitetsudnyttelse' og 'fuld strukturel beskæftigelse'. Alene valget af hvilke af efterspørgselskomponenterne, der vælges som tilpasningsvariabel vil have betydning for beregningen.

Men nok så alvorligt indebærer budgetlovens begrænsning på det maksimalt tilladte underskud på den strukturelle saldo til blot ½ pct. BNP (ca. 10 mia. kr.), at det i fremtiden – når en ny lavkonjunktur rammer – vil være umuligt at føre en ekspansiv finanspolitik. Christiansborgs politikere burde være taknemmelige for, at de ikke var pålagt denne restriktion i årene 2009-2012, se figur 1.

Omvendt bør det bemærkes, at der ikke i Budgetloven er en bremse, der kan hindre overophedning. Figur 1 illustrerer tydeligt, at ifølge finansministeriets beregninger var finanspolitikken i perioden 2004 til 2008 – uanset at der var et strukturelt overskud – for 'slap'. Budgetloven burde i stedet have et krav om, at saldoen for det strukturelle og det faktiske budget burde være (tilnærmelsesvist) sammenfaldende! Det ville bremse overophedning i en højkonjunktur, og tilsvarende bremse arbejdsløsheden i at brede sig i kølvandet på et eksogent efterspørgselsbortfald med efterfølgende lavkonjunktur i den private sektor.
Her ligger den pædagogiske opgave for finansministeriet (og for universitetsøkonomer), at forklare de politiske beslutningstagere, at der i 'gode tider' med overskud på budgettet ofte vil være brug for yderlig stramninger, mens der i en lavkonjunktur med fordel kan ekspanderes, idet de samfundsøkonomiske omkostninger begrænser sig til den øgede import fra udlandet.

Formand – og professor i økonomi, der i 2004 omtalte Stabilitetspagtens krav til de offentlige budgetter som 'stupido' – hvilket naturligvis pådrog ham en reprimande fra Den belgiske formand for Det europæiske Råd.

Figur



1. Den offentlige sektors strukturelle og faktiske saldo

Men nok så alvorligt betyder budgetlovens begrænsning på den tilladte størrelse af det beregnede underskud til max. ½ pct. af BNP (ca. 10 mia. kr.), at det i fremtiden vil være udelukket at føre en ekspansiv finanspolitik af nogen afgørende betydning.

Størrelsen heraf bestemmes i Finansministeriet, da den i sagen natur ikke kan fastlægges objektivt.
Kilde: Finansredegørelsen, 2014, kap. 7 s. 321, suppleret med nyere data

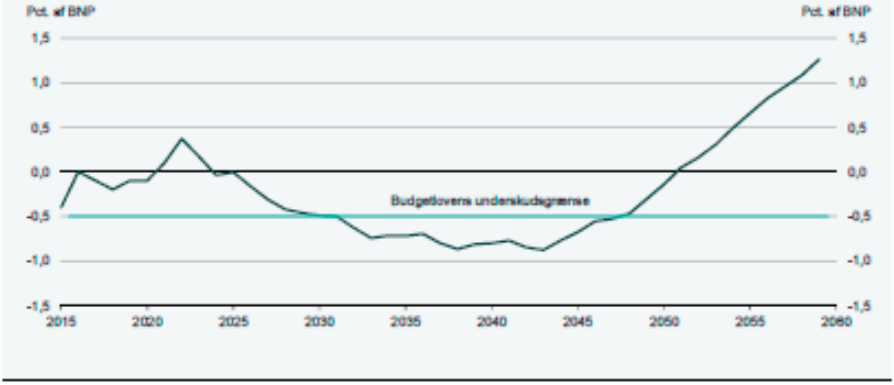
https://www.fm.dk/publikationer/2014/finansredegoerelse-2014

Men hvad nu hvis fremskrivningen udviser en overskridelse et stykke ude i fremtiden? Den ofte omtalte 'hængekøjeeffekt', der – ifølge finansministeriets modelberegninger – indebærer, at grænsen for det strukturelle underskud på ½ pct. af BNP vil overskrides i perioden fra 2030-2048, se figur 2 nedenfor. Bør der allerede nu tages forholdsregler?
Det oftest fremførte forslag er en yderlig og fremrykket hævning af pensionsalderen. Det vil øge arbejdsudbuddet, der i finansministeriets model er den 'hårde valuta'. Øget udbud af arbejdskraft føres ifølge modellen til en tilsvarende øget beskæftigelse, der aflaster de offentlige finanser.
De to argumenter tilsiger således, at nødvendigheden af at holde det strukturelle underskud under ½ pct. af BNP primært er en selvpålagt plage. For det første viser den endnu mere langsigtede fremskrivning, at de offentlige finanser er holdbare – der kommer på et senere tidspunkt et overskud, der fuldt ud kan dække den midlertidige demografisk betingede udgift. Desuden har EU indført den regel, at de lande der har

'solide' offentlige finanser godt i en periode kan have et strukturelt underskud op til 1 pct. af BNP – en grænse der ikke overskrides ifølge beregningen.
Men det vil kræve en ændring af Budgetloven – hvilket hermed være stærkt anbefalet!

Figur 1.3

Med uændret politik bliver de offentlige underskud for store omkring 2030-2050



Kilde: Danmarks Statistik og egne beregninger.

Kilde: Økonomi- og Indenrigsministeriet (2018), Danmarks konvergensprogram, april

3. Det realiserede offentlige saldo.

Stabilitets- og vækstpagten indebærer, at det offentlige underskud under normale omstændigheder højest må udgøre 3 pct. af BNP. Dette indebar, at den danske regering i 2013, som en konsekvens af overskridelsen i 2012, se figur 1 ovenfor, fik en henstilling fra EU-Kommissionen om at mindske budgetunderskuddet i 2013. Denne henstilling fulgte regeringen, dels ved at gennemføre besparelser, men nok så effektfuldt gennem 'operation budgetkosmetik'.

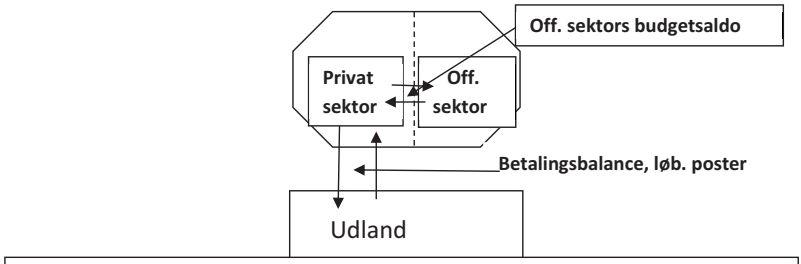
Budgetsminken bestod bl.a. i en fremrykning af beskatning af de hidtidige skattebegunstigede indbetalinger til kapitalpension. Disse indbetalinger kunne frigives til en ekstraordinær lav beskatning, som stort set alle benyttede sig af, idet pensionspengene kunne overflyttes til en på anden måde begunstiget pensionskonto i pengeinstitutter (og pensionskasser). Den ekstraordinære skatteindbetaling udgjorde ca. 100 mia.kr, hvilket forklarer den markante forbedring af det faktiske budget i 2013 og 2014, hvorved EU-Kommissionens henstilling blev tilgodeset!

Tilsvarende kan budgetsaldoen sminkes ved frasalg af offentlige reale aktiver og virksomheder.

Men det er direkte destabiliserende, at grænsen for underskuddets størrelse arbitrært er fastsat til 3 pct. af BNP – uanset arbejdsløshedens størrelse/konjunktursituationen.

Dette er en alvorlig mangel, idet det sætter snævre nedadgående grænser for de automatiske stabilisatorer. Grænsen burde – uanset diskussionen af behovet for at føre en aktiv finanspolitik, jfr. forrige afsnit – justeres efter konjunktursituationen både nedad; men også opad. Det var netop denne mangel på symmetri, der muliggjorde at flere af euro-landene i den gunstige konjunktur fra 2004-2007 vedblev med at have et underskud på budgettet. Her skulle grænsen har været tilpasset opad.
Det overses alt for ofte i den makroøkonomiske og politiske vurdering af budgetsaldoen, at den spejler en tilsvarende ubalance i den private sektor, hhv. opsparingsover- og underskud samt stillingen på betalingsbalancens løbende poster.

Figur 3. En simpel 3-sektor bogholderimodel



Bogholderimæssig sammenhæng:

Overskud privat sektor ≡ underskud off. sektor + betalingsbalanceoverskud over for udlandet (eksport ÷ import)

Pointen er, at i en relevant finanspolitisk analyse bør den offentlige budgetsaldo ikke analyseres isoleret; men altid som en del af en integreret makroøkonomisk model, hvor balancerestriktionen for 'the economy as a whole' respekteres. Det er kun under den ekstreme antagelse, at den private sektors opsparings/investeringsbalance sammenholdt med betalingsbalancen er i ligevægt, at det offentlige budget også vil være i ligevægt.

4. Udgiftslofter

Med budgetloven fra 2014 er der indført et krav om, at der skal aftales lofter for driftsudgifter (ekskl. ledighedsrelaterede) for staten, kommuner og regioner. Det er grundlæggende fornuftigt, idet planlægningshorisonten forlænges og regeringens politiske intention præciseres.

Betydningen heraf skal dog ikke overdrives, idet disse lofter løbende justeres i forbindelse med de årlige finanslovsforhandlinger, jfr. Den løbende diskussion af, om kravet om 2 pct. besparelse på en række ministeriers drift skal opretholdes.

Men også her har en asymmetrisk udformning af de sanktioner, der er knyttet til en evt. afvigelse fra det fastlagte loft vist sig uhensigtsmæssig. Kommuner og regioner pålægges en betydelig bøde, hvis de overskrider budgetloftet – uanset, at de har 'penge i kassen'. Omvendt er der ikke nogen mulighed for at overføre et evt. underforbrug til det næste driftsår. Det har i årene fra .2014 ført til at mindre-forbrug og dermed en strammere finanspolitik end den intenderede, hvilket har bidraget til at trække den økonomiske stagnation i langdrag.

Endelig burde der også have været en makroøkonomisk diskussion af, hvad der forstås ved en neutral udgiftspolitik/finanspolitik. I og med at skatteindtægterne ved uændrede skattesatser stort set følger udviklingen i BNP, vil en konjunktur-neutral finanspolitik indebære; at de offentlige udgifter udgør en konstant andel af BNP, hvilket både indebærer en stigning i reallønnen for offentligt ansatte, men også en stigning i øvrige driftsudgifter, herunder velstandsregulering af overførselsindkomster.

5. Budgetloven – giver den mening?

'Budgetloven blev vedtaget af et bredt flertal i Folketinget i maj 2012 og har i praksis haft virkning fra finansåret 2014'.

På baggrund af ovenstående gennemgang af de samfundsøkonomiske implikationer og konsekvenser af Budgetloven, vil jeg tillade mig at have mine tvivl, om det 'brede flertal i folketinget' havde gjort sig klart, at de hermed i praksis afskrev sig muligheden for at føre en aktiv finanspolitik i fremtiden.

Budgetlovens regelsæt indeholder ydermere den uhensigtsmæssighed, at det vil tendere mod at forlænge en lavkonjunktur og ikke indeholder nogen bremse på en højkonjunktur. Det betyder, at med Budgetloven er risikoen for en gentagelse af forløbet fra 2004-2018 blevet institutionaliseret. Det har allerede manifesteret sig i det såkaldte finanspolitiske råderum, der ved indgangen til en opgangskonjunktur blev opgjort til 35 mia. kr. (ca. 2 pct. af BNP) og som politikerne er begyndt at omsætte til ikke-finansierede finanspolitiske initiativer – uden hensyntagen til konjunktursituationen. Når lavkonjunkturen sætter ind – jfr. Multiplikator-accelerator teorien – så vil budgetlovens krav om balance på det strukturelle budget hindre, at samfundsøkonomisk rationelle aktiviteter sættes i gang. Ligesom overskridelse af 3 pct. grænsen vil føre til offentlige besparelser og flere udbudsreformer!

Kilder:

De økonomiske Råd (2018), Dansk økonomi, efteråret. København: eget forlag

Greve, B. og J. Jespersen (red.), (2019), Den offentlige sektors økonomi, 4. udg. København: Hans Reitzels Forlag

De økonomiske råds sekretariat (2015), https://dors.dk/offentlige-finanser/finanspolitiske-begreber

https://www.dors.dk/offentlige-finanser/finanspolitiske-begreber/finanspolitisk-holdbarhed

https://www.fm.dk/oekonomi-og-tal/budgetlov-og-udgiftslofter

https://www.fm.dk/oekonomi-og-tal/oekonomisk-analyse/2017/holdbarhed-og-generationsmaessig-neutralitet

https://www.skm.dk/skattetal/statistik/generel-skattestatistik/kapitalpensionsafgift-og-rabatordning-i-2013,-2014-og-2015

Jespersen, J. (2007), Makroøkonomisk Metodologi, København: Jurist- og økonomforbundets Forlag

Jespersen, J. (2018), Finansministeriets økonomiske råderum: luftkastel eller realitet? i 'Symposium i anvendt statistik, København: Det Nationale Forskningscenter for Arbejdsmiljø, s.219-27

Kokholm, Th. (2018), Topøkonomer anbefalede budgetloven: Nu undsiger de den. Danske Kommuner, 19. februar.

Økonomi- og Indenrigsministeriet (2018), Danmarks konvergensprogram, april

Hawthorne eller Tinto?

En analyse af førsteårsfrafaldet ved de erhvervsøkonomiske bacheloruddannelser

Af

Mogens Dilling-Hansen

Department of Economics and Business Economics

Aarhus University

Mail: dilling@econ.au.dk

Resumé

”*I denne analyse præsenteres modeller for frafaldet på de videregående uddannelser. Modellerne er meget ambitiøse og dermed svære at estimere; men de foreløbige resultater fra et pilotprojekt ved de erhvervsøkonomiske uddannelser på Aarhus Universitet tyder på, at der faktisk er en række simple tiltag, der i hvert fald på det korte sigt mindsker førsteårsfrafaldet.*”

1 Indledning

Specielt på de erhvervsrettede og videregående uddannelser er frafaldet på de videregående uddannelser i Danmark et stort problem. Det fremgår af Uddannelses- og Forskningsministeriets seneste analyse præsenteret i Styrelsen for Forskning og Uddannelse (2018), hvor både frafaldet og studieskifte er beskrevet. I Uddannelses- og Forskningsministeriet (2018) er frafaldet blandt akademiske bachelorer opdelt efter studieskifte eller ej: I 2010 var frafaldet på 6.048 studerende ud af 2006-populationen på 19.807 bachelorer og tallene er steget til 10.044 frafaldne ud af 28.850 studerende i 2017. Andelen af frafaldne, der afbryder uden studieskifte, er i samme periode steget fra 75% til 81%.

Er det godt eller skidt at omkring 40% af akademiske bachelorstuderende på universiteterne ikke gennemfører den uddannelse, som de er gået i gang med? Formålet med denne analyse er at komme lidt tættere på en forståelse af frafaldet og derudover at analysere effekten af et konkret frafaldsprojekt ved de erhvervsøkonomiske bacheloruddannelser på Aarhus Universitet i 2018.

Opbygningen af artiklen er følgende. Kapitel 2 diskuterer frafaldsproblemet generelt og der fortsættes i kapitel 3 med en præsentation af Tintos model for fraflad. Kapitel 4 præsenterer frafaldsprojektet og resultaterne af analyserne præsenteres i kapitel 5. Kapitel 6 diskuterer rækkevidden af resultaterne og der peges på behovet for yderligere viden og forskning på området.

2 Frafaldsbegrebet

Frafald er defineret som en bogholderi-baseret opgørelse af en kohorte af studerende:

$$(1) \quad Bestand_t = Bestand_{t-1} + Frafald_t$$

hvor tiden t udtrykker den relative tid for kohorten og $Bestand_{t=0}$ er det faktiske antal optagene studerende de pågældende år.

I nedenstående tabel 1 er frafaldet opgjort opdelt efter år og studiestart, og de ses at langt det største frafald sker det første studieår: Førsteårsfrafaldet er i 2013-kohorten på 17% og i perioden 2013-17 (fire studieår) er det samlede frafald på 35%.

Der startede 28.850 akademiske bachelorer ved universiteterne i 2013, og efter det 4. studieår er der kun omkring 18.750 studerende der har fuldendt (eller som har mulighed for at fuldende) deres påbegyndte uddannelse.

Tabel 1 Første til fjerdeårsfrafald for akademiske bachelorer på universiteterne, 2013-17

	Årgang 2012	Årgang 2013	Årgang 2014	Årgang 2015	Årgang 2016
Frafald i løbet af 1. studieår	16	17	18	17	17
Frafald i løbet af 1. og 2. studieår	25	28	29	27	
Frafald i løbet af 1., 2. og 3. studieår	30	33	34		
Frafald i løbet af 1., 2., 3. og 4. studieår	33	35			

Noter: Kilde er Uddannelses- og Forskningsministeriet og yderligere detaljer kan ses i Uddannelses- og Forskningsministeriet (2018). Tallene er opgjort i procent og en studieårgang dækker over alle studerende, som er påbegyndt en akademisk bacheloruddannelse på et universitet i studiestartsåret.

Blandt de studerende, der afbryder uddannelsen, skifter omkring 75-80% af gruppen til et andet studium, medens de resterende 20-25% har afbrudt uddannelsen uden studieskift.

Det er med andre ord omkring halvdelen af det samlede frafald, der sker det første studieår, og derfor er der naturligt mest fokus på dette frafald. Medens frafald på et senere tidspunkt typisk tilskrives en social begivenhed, så er der typisk flere forklaringer på frafaldet det første studieår: I Styrelsen for Forskning og Uddannelse (2018) er der baseret på en spørgeskemaundersøgelse opgjort andelen af studerende, hvor betydningen har været stor (konkret svarene 4-5 på en skala fra 1til 5, og top-10 er (i) "Uddannelsens faglige niveau" (45%), (ii) "De jobs, som uddannelsen førte til, var ikke noget for mig", (iii) "Uddannelsen var alt for teoretisk", (iv) "Jeg følte mig ensom på studiet", (v) "Kvaliteten af undervisningen", (vi) "Det sociale miljø på uddannelsen var for dårligt", (vii) "Jeg havde ikke et fagligt samarbejde med min medstuderende", (viii) "Kontakten med underviserne", (ix) "Personlige forhold uden for studiet" og "Jeg følte mig ikke velkommen på studiet" (29%).

På de erhvervsøkonomiske uddannelser ved Aarhus Universitet er førsteårsfrafaldet højere end landsgennemsnittet, og desuden svinger frafaldet op til 5% point fra et år til det næste.

3 Tintos model for frafald

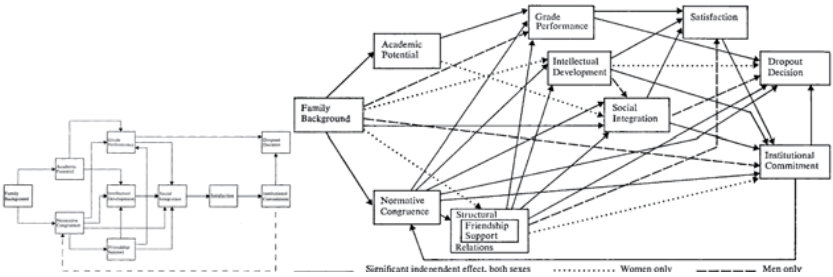
Tinto er ikke den eneste, der har kigget på frafald fra videregående uddannelser. Den empiriske del er vokset i teknisk henseende, og således præsenterer Cabrera et. al. (1993) modeller estimeret som strukturelle modeller. Fælles for modellerne - se fx Spady (1970, 1971), Tinto (1975), Bean (1980) og Cabrera (1993) – er at de er indeholder forklarende elementer i fire grupper:

i. Akademisk potentiale
ii. Integration – social såvel som akademisk (påvirkes bl.a. af "friendship support")
iii. Commitment – både individuelt og institutionelt
iv. Personlige egenskaber, tilfredshed mm.

Resultaterne er noget svingende med nogle markante ekstremer i form af en total forkastelse af alle "Tinto-inspirerede" hypoteser, se fx Brunsden et. al. (2000). Det generelle billede er dog, at der findes den forventede positive sammenhæng mellem de fire typer af forklarende variable og generel performance på studiet, dvs. "ikke-dropout decision".

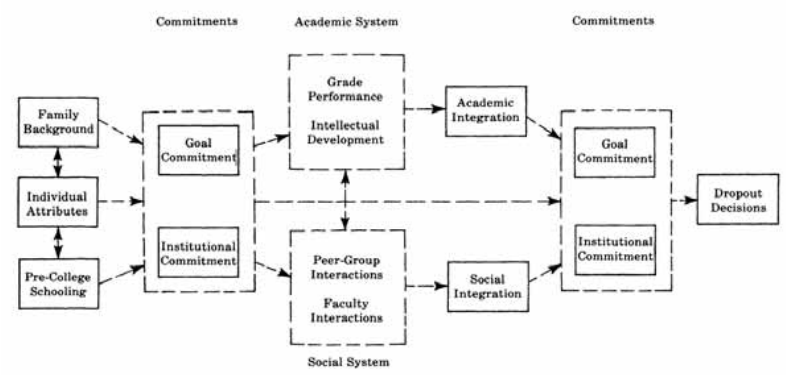
Figur 1 og 2 viser nogle eksampler på fulde strukturelle modeller til forklaring af "dropout decision", og specielt Spady's empiriske version 2 i figur 1 (figur til højre) viser modellens signifikante relationer. Analyserne er baseret på 683 observationer af førsteårs studerende ved University of Chicago helt tilbage i 1965, men resultaterne er repræsentative for de efterfølgende empiriske analyser.

Figur 1 En model for "Dropout Process" – eksempel fra Spady (1970)



Noter: To eksempler på empiriske modeller for "Dropout Decision"; kilder er hhv. Spady (1970) og Spady (1971)

Figur 2 Tinto's model for "Dropout Process"



Kilde: Tinto (1975)

Der er særligt tre fundamentale problemer ved disse fulde modeller. *For det første* er det særdeles vanskeligt at påvise robuste kausale sammenhænge, og det viser sig bl.a. ved at der øjensynligt ikke er en sammenhæng mellem generel tilfredshed og beslutningen om dropout, jf. figur 1.

For det andet er der i alle frafaldsmodeller mange dimensioner, som skaber et behov for mange – og valide - data. Registerdata kan bruges til at analysere visse "hårde sammenhænge" (fx demografi og karakterer); det må dog generelt konstateres, at normative dimensioner samt begreber som "potentiale", "udvikling" og "integration" nok kan måles fx gennem spørgeskemabaserede undersøgelser, men sikring af validitet er ganske vanskeligt. *For det tredje* er modellerne basalt set sekventielle, og specielt måling af frafaldet (dropout) typisk en måling, der kan foretages på et senere tidspunkt end tidspunktet, hvor forklarende variable måles.

Spørgsmålet om anvendelse af frafaldsmodeller diskuteres i bla. Tinto (2006). På den ene side er en fuld analyse baseret på Tintos model meget ambitiøs og omfattende, men til gengæld er analyserne nyttige til forståelse af det komplicerede samspil mellem personlige egenskaber, karakterniveau og institutionens indsats for at bevare/højne kvaliteten af undervisningen.

På den anden side er alle tiltag, som forbedrer chancerne for at forhindre frafald – bedre optagelsesprocedurer, bedre undervisning, bedre social og faglig integration – ikke vigtige, hvis den studerende i mellemtider er droppet ud. Med andre ord argumenter Tinto (2006) for at der indføres nogle meget simple procedurer, der forlænger den effektive studietid baseret på den formodning/teori, at jo længere tid den studerende "opdrages" jo større er sandsynligheden for at undgå frafald. I denne analyse er denne hypotese kan Hawthorne-effekten, og den er opkaldt efter de klassiske eksperimenter, som blev foretaget i perioden 1924-32: Elton Mayo identificerede positive effekter også i kontrol-gruppen, og tolkningen er den, at der alene ved at der udvises interesse for det enkelte individ sker en produktivitetsforøgelse. Eksperimenter, hvor der skrues op for lyset, giver den forventede stigning i produktiviteten; men effekten af at der skrues ned for lyset er tilsvarende positiv, fordi den observerede objekt gennem observationer følte sig betydningsfuld.

Er det virkelig så simpelt? Kan frafaldet reduceres blot ved at forlænge den tid i starten af studiet, hvor den studerende stadig er i den afklarende fase?

Det frafaldsprojekt, der beskrives i næste kapitel, er et eksempel på et projekt, hvor der gennem målrettet indsats i det første semester af studiet forsøges at reducere førsteårsfrafaldet.

4 Frafaldsprojektet i 2018: Erhvervsøkonomiske bachelorer ved AU

I perioden august 2018 til december 2018 er der gennemført et pilotprojekt, der har til formål at reducere frafaldet, gennem en målrettet indsats for at forlænge den tid, hvor den studerende faktisk er studieaktiv. Formelt set er der tale om et pilotprojekt, som Det erhvervsøkonomiske Studienævn er ansvarlig for, og projektet har modtaget finansiel støtte fra Aarhus Universitets strategiske midler. De studier, der deltager i projektet, er HA- og BSc-studierne i hhv. Aarhus og Herning og BA(SOC)-studiet, som alene udbydes i Aarhus. Resultaterne, der præsenteres her, er alene fra uddannelsesstedet Aarhus, dels for at sikre sammenlignelige tal og dels fordi data fra Herning endnu ikke har gennemgået validitetscheck. Samlet er der tale om 841 (521+193+66 +61) studerende med uddannelsessted i Aarhus. De studerende, der deltager i projektet, er alle optaget september 2018, og de følger alle fire fag på det erhvervsøkonomiske bachelor-studiums første semester: Organisation, matematik, statistik og eksternt årsregnskab.

Der er ikke tale om et eksperimentelt design, så den endelige effekt måles ved at sammenligne førsteårsfrafaldet målt i 2019 med frafaldet i 2017 (knap 20%) og 2018 (knap 25%). Der ændres heller ikke på den grundlæggende forelæsningsform suppleret med øvelseshold; de fire forelæsningshold har typisk 150-200 deltagende studerende, medens de 31 instruktorhold typisk har 20-30 studerende på holdlisterne.

Pilotprojektet består at to dele: Et gruppedannelsesprojekt og en mentorordning.

4.1 Om gruppedannelsesprojektet

Formålet med denne del af pilotprojektet er at inddele alle studerende i grupper, således der skabes opmærksomhed på den enkelte studerendes tilknytning til institutionen og til de øvrige studerende. Formålet er dermed at skabe gruppedynamik ved at styrke de faglige og sociale bånd på de enkelte øvelseshold.

Grupper består at fire studerende, og de dannes allerede inden studiestart: På baggrund af et spørgeskema besvaret i rus-ugen (ugen før studiestart) har alle studerende modtaget besked om deres arbejdsgruppes sammensætning. Det har været frivilligt om den enkelte studerende deltager i gruppe-projektet; men det kræves at spørgeskemaet er besvaret, hvis man vil deltage – ca. 10 studerende ud af 800 studerende har aktivt fravalgt gruppe-ordningen.

4.1.1 Procedure for gruppeinddeling

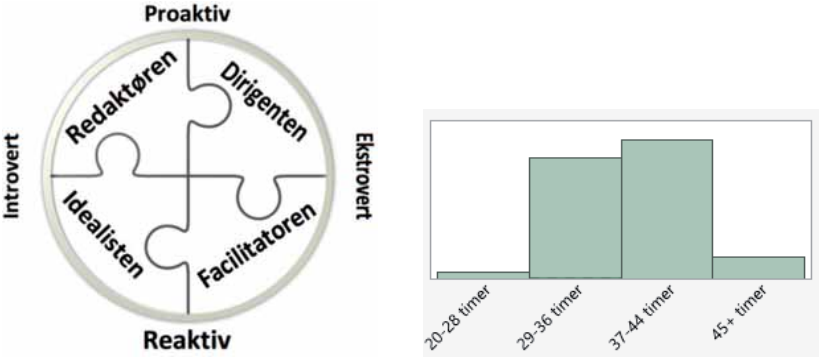
Med udgangspunkt i Tinto's model ville det være oplagt at inddele studerende efter både faglige og sociale ressourcer. Private firmaer tilbyder allerede den slags ydelser med kortlægning af personlige egenskaber som grundlag for at matche studerende i funktionelle grupper. Denne procedure er imidlertid ret kostbar (typisk med samlede omkostninger omkring 1.000 kr per studerende), og derfor afventer denne tilgang dokumentation af frafaldet (læs: ekstra STÅ-indtægter). Det andet ekstrem ville være at bruge Hawthorne effekten som begrundelse for at inddele de studerende efter et mere eller mindre tilfældigt princip: Studerende kunne vælge selv eller grupperne dannes ved tilfældig lodtrækning.

4.1.2 Gruppeinddeling efter personlige kompetencer

I pilotprojektet er valgt en procedure midt mellem Tinto og Hawthorne principperne i form af, at grupperne er dannet på baggrund af den studerendes selvrapporterede kompetenceprofil i stil

med Belbin tests, se fx Belbin et.al. (1979) og Prichard & Stanton (1999). Figur 3 viser den principielle opdeling af studerende efter fire typer af kompetencer: I forhold til de øvrige studerende kan en studerende være overvejende proaktiv, reaktiv, introvert hhv. ekstrovert, og grupperne er dannet således der er én studerende med overvejende styrke inden for én af de fire kompetencer.

Figur 3 Studerendes kompetenceprofil, "rollehjul", og planlagt ugentlig arbejdsindsats



Proceduren er baseret på et traditionelt Likert-skaleret batteri af spørgsmål, som søger at afdække de fire typer af kompetencer; her eksemplificeret med spørgsmål om den studerende evne til få gruppearbejdet til at fungere:

Hvor enig eller uenig er du i følgende udsagn:

	Meget uenig	Uenig	Hverken eller	Enig	Meget enig
Jeg er god til at motivere mine gruppemedlemmer til at bidrage til arbejdet.	○	○	○	○	○
Selvom jeg som udgangspunkt er interesseret i alles synspunkter, kan jeg hurtigt tage et standpunkt, når en beslutning skal træffes.	○	○	○	○	○
Jeg er god til at fordele arbejdsopgaverne mellem gruppens medlemmer med det formål at optimere processen.	○	○	○	○	○

Data er checket med anvendelse af eksplorativ faktoranalyse, og kompetencerne er efterfølgende beregnet som simple gennemsnit af de tre manifeste variable.

De endelige grupper er dannet på baggrund af den samlede kompetenceprofil kombineret med en subjektiv vurdering af den studerendes ambitionsniveau.

4.2 Om mentorprojektet

Denne del af projektet afvikles i løbet af det første semester, og den grundlæggende tankegang er, at der ofte ved studiestart er behov for at modtage gode råd og vejledning af ældre studerende, der har prøvet at stå i samme situation, Den ældre studerende – mentoren – er ansat efter ansøgning, hvor der dels er lagt vægt på at mentoren faktisk har bestået fagene på det første semester og dels at mentoren har vist interesse for at gøre en forskel på det sociale plan.

Der er ansat én mentor per øvelseshold, og arbejdsopgaverne for de 31 mentorer er kun styret i meget generelle termer: Det overordnede formål er at sørge for at grupperne udnytter den dynamik, der kan opstå ved velfungerende grupper; men også individuelle (sociale) udfordringer kan tages op. Der afholdes mindst tre møder i løbet af semestret og der skal foretages simpel registrering af fremmøde samt aktivitet generelt. Den studiemæssige status for den enkelte studerende er opgjort medio december 2018, hvor status for den studerende er opgjort efter følgende princip:

Tabel 2 Inddeling af studerende efter studiestatus, december 2018

Status	Beskrivelse
A	Den studerende er aktiv i den oprindelige gruppe (335 studerende; 42%)
A1	Den studerende er aktiv i en let modificeret gruppe (187 ; 23%)
B	Den studerende er aktiv i en ny gruppe (197 ; 24%)
C	Den studerende er formelt aktiv, men følger ikke undervisningen (18 ; 2%)
D	Den studerende er ikke aktiv og har muligvis slet ikke været til stede (28 ; 3%)
D1	Den studerende er droppet ud i løbet af semestret (41 ; 5%)

Tallene i parentes i tabel 2 er det faktiske antal målt i december 2018, og det ses at 8% på det tidspunkt reelt er droppet ud. Det formelle antal dropout er til gengæld meget begrænset.

5 Resultater

Den endelige vurdering af pilotprojektet vil i tråd med andre overlevelsesanalyser først kunne foretages efter det første studieår, hvor førsteårsfrafaldet kan sammenlignes med tidligere tal. Analyserne, der præsenteres i dette afsnit, er dels en analyse af det faktiske frafald konstateret

december 2018, altså inden den første eksamen afvikles, og dels en analyse af de studerendes vurdering af frafaldsprojektet. Kilderne til de to analyser er den spørgeskemabaserede kompetenceanalyse, der blev foretaget før studiestart (august 2018) og den afsluttende spørgeskemaundersøgelse af pilotprojektet foretaget december 2018. Denne sidste evaluering skal naturligvis vurderes i lyset af, at det kun er studerende, der stadig er aktive på studiet, som har svaret på spørgeskemaet, og antallet af observationer falder dermed fra 841 deltagende studerende til 233 deltagere i december-undersøgelsen.

5.1 Frafaldsanalyse

Det fremgår af tabel 2, at 87 studerende reelt er droppet ud af studiet i løbet af de første tre måneder af studiet. Til analyse af sandsynligheden for et frafald er der anvendt logistisk regression med den binære variabel "frafald" som afhængig variabel. Imidlertid er den samlede andel af frafaldne kun 10% og derfor er metoden ikke særlig velegnet set ud fra et statistisk synspunkt: Prediktionsevnen bliver lav og modellen kan reelt kun forudsige, hvad frekvenstabellen siger, nemlig at hvis en studerende trækkes ud tilfældigt, så er det mest sandsynligt at det er en aktiv studerende.

Tabel 3 viser en simpel logistisk model for sandsynligheden for et dropout. Køn er ikke signifikant, og det gælder også for de personlige kompetencer: Det betyder ikke noget for sandsynligheden for et frafald om den studerende besidder specielle kompetencer.

Tabel 3 Sandsynligheden for studieaktivitet

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	12,05679	2	24,11358	<,0001*
Full	263,20354			
Reduced	275,26033			

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	1,400479	0,1887159	55,07	<,0001*
BSC	0,65902833	0,3081504	4,57	0,0325*
HA	1,30541828	0,264626	24,34	<,0001*

For log odds of Aktiv/Dropout

Confusion Matrix

Training

Actual Frafald	Predicted Count Aktiv	Dropout
Aktiv	755	0
Dropout	85	0

Noter: Metode er logistisk regression med frafald som afhængig variabel (bemærk at kun 10% er registreret i december 2018 som frafaldne). JMP er anvendt i analyserne, hvorfor "confusion matrix" er modellens prediktionsevne. Parametrene beskriver sandsynligheden for at være aktiv, P(aktiv).

Tolkningen af tabel 3 er som forventet. Den estimerede model er signifikant, men dens evne til at fremskrive er begrænset til at kunne forudsige at den typiske studerende stadig er aktiv. De to forklarende variable, HA og BSC, er dummies for de to store uddannelser, HA-studiet og den tilsvarende engelsksprogede uddannelse. Tolkningen skal dog foretages med varsomhed, fordi frafaldet fordelt på uddannelsesretninger også dækker over at der er oprettet 31 øvelseshold, og der er inden for de enkelte øvelseshold meget store variationer i antallet af aktive studerende: 18 ud af de 31 øvelseshold har et frafald på under 10%, men der er 3 hold med et frafald på mindst 20%. Der er noget, der tyder på at velfungerende øvelseshold har afgørende betydning for at den enkelte studerende fortsat er studieaktiv.

5.2 Tilfredshedsundersøgelse

Resultaterne i dette afsnit er baseret på de studerendes feedback på pilotprojektet, og undersøgelsen blev afviklet i december 2018. I denne præsentation er der anvendt 233 observationer.

Tabel 4 Tilfredshed med studiegrupper og mentorordning

Studiegrupper	***Gennemsnit***	***Spredning***
Hjælp til undervisning	5,0	1,7
Hjælp til at forstå pensum	5,0	1,6
Hjælp til faglig integration	5,2	1,6
Hjælp til social integration	5,1	1,6
Samlet vurdering	5,7	1,8
Mentorordning	***Gennemsnit***	***Spredning***
Generel vurdering	4,3	1,9
Tvivl om mentorens rolle	4,3	1,9

Noter: Der måles på en 7-trins Likert-skala

Tabel 4 er en summarisk beskrivelse af den aktive studerendes vurdering af pilotprojektets to komponenter, gruppedannelse og mentorordningen. Ikke overraskende er den samlede vurdering meget positiv; men det ikke enkelte konkrete punkter der skaber tilfredsheden. Mentorordningen er også vurderet positivt, men dog på et noget lavere niveau og med en større spredning i svarene. Den større spredning dækker over en række kvalitative vurderinger, som svinger tilsvarende og dækker over det forhold, at nogle studerende virkelig har haft personligt udbytte af at

have "nogle brede skuldre at læne sig op ad" – til gengæld har andre ikke set et konkret udbytte af mentorerne når problemer med gruppedannelserne skulle løses.

Mentorordningen er løntung og den største udgift på budgettet for pilotprojektet, så ud fra et økonomisk perspektiv er det gruppedannelsesprojektet, der giver størst værdi for pengene.

Tabel 5 Tilfredshed med studiet

Tilfredshed med studiet	5,5	1,3
Generel trivsel	5,6	1,4
Følelse af stress	5,0	1,6
Svær overgang fra gymnasiet	4,4	1,7

Noter: Der måles på en 7-trins Likert-skala

Tabel 5 viser et summarisk overblik over de studerendes vurdering af studiemiljøet. Det gennemgående træk er, at de aktive studerende er tilfredse med studiet; men det kan også bemærkes at følelsen af at være stresset er fremherskende.

Det er generelt svært at forklarende den generelle tilfredshed med studiet, medens følelsen af at være stresset har to klare forklarende faktorer: Ved anvendelse af simpel regressionsanalyse kan det påvises, at mænd generelt føler sig mindre stressede end kvinder og studerende med matematik på A-niveau på tilsvarende måde føler sig mindre stressede.

Tabel 6 Tilfredshed med fagligt indhold af studiet

	Årsregnskab	*Organisation*	*Matematik*	*Statistik*
Sværhedsgrad af fag	5,1	3,8	5,0	5,6
... for studerende med mat A	5,2	3,9	4,2	5,4
Interessant fag	4,8	5,0	4,7	4,3
... for studerende med mat A	4,8	4,9	5,2	4,2
Betalt ekstraundervisning	13%	6%	35%	27%

Noter: Der måles på en skala fra 1 til 7. Procent-tallene i sidste linie angiver den andel af studerende, der har betalt for ekstraundervisning i løbet af semestret.

Endelig er der i tabel 6 vist hvorledes de studerende vurderer de fag, som indgår i det første semester af bachelorstudiet. Hvis et gennemsnit på fire udtrykker en form for neutralitet, så er betrykkende at selv redskabsfag vurderes pænt med hensyn til om fagene er interessante. At studerende med matematik på A-niveau i den adgangsgivende eksamen vurderer sværhedsgraden anderledes end de øvrige studerende med matematik på B-niveau er måske heller ikke så overraskende; men til gengæld må det bekymre at op mod hver tredie studerende føler det er nødvendigt at købe sig til ekstra undervisning.

6 Afrunding

Et væsentligt formål med pilotprojektet er at blive klogere. Hvad er det egentligt, der foregår på det første studieår? Hvornår falder de studerende fra og hvorfor gør de det?

Formålet med denne analyse er at præsentere de foreløbige resultater. Set ud fra et generelt perspektiv er det godt at konstatere, at der ikke på kort sigt er et kønsspecifikke frafaldsproblem, ligesom de to store uddannelser tilsyneladende fungerer godt.

De foreløbige resultater tyder på, at gruppedannelser faktisk har en positiv effekt på frafaldet på det korte sigt. Til gengæld må tiden vise, om frafaldsprojektet faktisk har mindsket frafaldet eller blot udskudt det!

Referencer

Bean, J. P. (1980), Dropouts and Turnover: The Synthesis and Test of a Causal Model of Student Attrition, *Research in Higher Education,* 1980, vol. 12, no. 2, 1980, pp. 155-187.

Belbin, R.M., B. R. Aston & R. D. Mottram (1979)**,** Building Effective Management Teams, *Journal of General Management,* March 1976.

Brunsden, V, M. Davies, M. Shevlin & M. Braecken (2000), Why do HE Students Drop Out? A test of Tinto's model, *Journal of Further and Higher Education*, 24:3, pp. 301-310.

Cabrera A.F., A. Nora & M.B. Castañeda (1993), College Persistence. Structural Equations Modeling Test of an Integrated Model of Student Retention, *Journal of Higher Education*, vol. 64, No. 2, pp. 123-139.

Malhotra, N. K., D. F. Birks & P. Wills (2012), *Marketing Research. An applied approach*, Pearson Education Limited, Essex, 2012

Prichard, J & N. A. Stanton (1999), Testing Belbin's team role theory of effective groups, *Journal of Management Development* 18(8), pp. 652-665

Spady, W. G. (1970), Dropouts from Higher Education: An Interdisciplinary review and synthesis, *Interchange,* pp. 64-85.

Spady, W. G. (1971), Dropouts from Higher Education: Toward an Empirical Model, *Interchange,* vol. 2 (3), pp. 38-62.

Styrelsen for Forskning og Uddannelse (2018), "*Frafald og Studieskifte*", marts 2018, 53 sider, web-link: https://ufm.dk/publikationer/2018/filer/frafald-og-studieskift.pdf

Tinto, V. (1975), Dropout from Higher Education: A Theoretical Synthesis of Recent Research, *Review of Educational Research,* vol. 45, no. 1, pp. 89-125.

Tinto, V. (2006), Research and Practice of Student Retention: What next?, *Journal of College Student Retention,* vol. 8(1), pp. 1-19.

Uddannelses- og Forskningsministeriet, 2018, "*Frafald blandt akademiske bachelorstuderende fra universiteterne inden for 4. år efter studiestart, jf. Danske Universiteters opgørelse"*, notat, web-link: https://ufm.dk/uddannelse/statistik-og-analyser/frafald/notat-om-frafald-og-studieskift.pdf

SAS® Software Certified Young Professionals

Expand your opportunities with SAS® skills
and become an analytics expert

1 Access SAS Software + 2 Learn SAS + 3 Prepare for Certification + 4 Take the Exam

Free SAS software for academic, noncommercial use. An interactive, online community. Free access to superior training and documentation. Free registration for the certification exam. Gain the analytical skills you need to secure your future.

SAS® SCYP
SAS® SOFTWARE CERTIFIED YOUNG PROFESSIONALS

www.sas.com/scyp

The SAS SCYP Program is offered to bachelor and master students of universities and university colleges in countries where we organize exam testing events and job-events for students. Check out the registration form on www.sas.com/scyp to see if the SAS SCYP Program is offered in your country. You do not need to have prior knowledge about SAS before starting on the SAS SCYP Program.