A Vine Copula Panel Model for Day-Ahead Electricity Prices

Valberg-Madsen, Janus Sejersbøll; Høg, Esben; Christensen, Troels Sønderby; Pircalabu, Anca

Published in:
Symposium i Anvendt Statistik 2020

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

Symposium i anvendt statistik 2020

Medarrangører:
Institut for Økonomi, Aarhus BSS, Aarhus Universitet
Det Nationale Forskningscenter for Arbejdsmiljø

SYMPOSIUM I ANVENDT STATISTIK

27.-28. januar 2020

Redigeret af Peter Linde
på vegne af organisationskomiteen for
Symposium i Anvendt Statistik

Støttet af SAS Institute Inc.

Institut for Økonomi, Aarhus BSS, Aarhus 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 Institut for Økonomi, Aarhus BSS, Aarhus Universitet vært for symposiet, hvilket vi gerne vil takke for. Symposiet arrangeres af Symposium i Anvendt Statistik og Det Nationale Forskningscenter for Arbejdsmiljø og Økonomisk Institut, Københavns Universitet er medarrangør. Den faglige forening Symposium i Anvendt Statistik er ansvarlig for det faglige program og økonomien.

Denne publikation indeholder foredragene fra det 42. 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 978-87-989370-0-5

Trykt hos PRinfoTrekroner i 150 eksemplarer

Organisationskomiteen for Symposium i Anvendt Statistik 2020

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, CBS
Solbjerg Plads 3
2000 Frederiksberg
stgg@cbs.dk

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

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

Sören Möller
Faculty of Health Sciences
Syddansk Universitet
J. B. Winsløws Vej 19
5000 Odense C
Moeller@health.sdu.dk

Indholdsfortegnelse

Statistiske metoder

Anvendt statistik og metode

Økonomi og samfund

Mikkeller and Weather Data

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), la Cour et al (2018) and la Cour et al (2019) 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) and la Cour et al (2019), we focus mainly on investigating the predictive power of Facebook data (FB) with respect to sales. However in the present case we will also add weather data to try and increase the accuracy of our models and predictions. 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 have decided to change the frequency of the data from daily to weekly. Such an operation will remove some of the noise in the data and still leave us with enough observations to run our time series models. We will still try various transformations of the social media variables and also add dummies for specific types of weather events. 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).

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". When aggregating to the weekly level our total sample contains 144 observations with the last 13 observations left for an out-of-sample forecasting exercise.

Prior to analysis we index the daily 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.

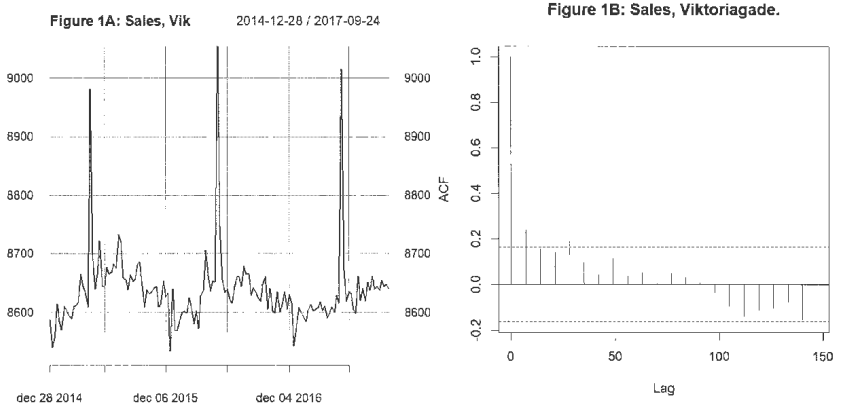
Finally we have tried more manually to search for specific words in the post from the FB administrators of both Mikkeller headquarters and Mikkeller Viktoriagade that

relate to the Mikkeller Beer Celebration Copenhagen (MBCC), an event that takes place every year in April or May and that attracts a lot of extra customers also from abroad and therefore leads to larger sales figures for those specific weeks.

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. We replaced these daily values with the average values of the series and the performed the aggregation to the weekly frequency using these imputed values.

We show graphs of our main series of interest and its ACF in figures 1A and 1B.



The weekly 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 MBCC weeks. In this respect the pattern looks quite similar to the one from the daily series although less erratic, see laCour et al (2019). In figure 1B, we see a significant spike at lag 1 and also one at lag 4. The lag 4 spike may indicate some monthly seasonality. The autocorrelations are not large enough (all less than 0.30 in absolute size) to make us suspect non-stationarity of the series. Recall that we have pre-processed the sales data to anonymize them. 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 The regression models

We use dynamic regression models (or ADL/ARX models). With weekly data of beer sales in a bar, we expect some seasonality. In the present modelling we will try to capture the seasonality by using lags of the sales variable and also by lags of the FB variables. Furthermore, the inclusion of weather variables might help capture the seasonal patterns. 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. Then we extend the model by adding blocks of regressors to capture holidays and the Mikkeller Beer Celebration Copenhagen event. We also experiment with using weather data instead of the monthly dummies to capture the seasonality. The final block of regressors we add to the model contains the first four lags of the each FB variable: the posts by the bar, the posts by the head quarters, the likes for the posts of the bar and the likes for the posts of the head quarters.

The model we take as the starting point for the benchmark model is:

(1) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_4 y_{t-4}$ + trend + monthly dummies + ε_t $\quad t = 1, .., T$
(GUM1)

where y is sales. The error term, ε_t, is assumed to fulfill the standard assumptions for OLS estimation. These assumptions hold true for all the model equations.

The next model has model (1) as the starting point and then regressors for holidays and the MBCC events are added:

(2) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_4 y_{t-4}$ + trend + monthly dummies
+ holidays + MBCC dummies + ε_t $\quad t = 1, .., T$
(GUM2)

Next, we replace the monthly dummies by the block of weather variables to see if adding such variables can do the same or an even better job than the dummies:

(3) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_4 y_{t-4}$ + trend + weather variables
+ holidays + MBCC dummies + ε_t $\quad t = 1, .., T$
(GUM3)

Finally, we extend model (3) with the FB variables block:

(4) $y_t = \beta_0 + \gamma_1 y_{t-1} + \ldots + \gamma_4 y_{t-4}$ + trend + weather variables
+ holidays + MBCC dummies + lagged FB variables + ε_t $\quad t = 1, .., T$
(GUM4)

We call all our model 'GUM' followed by the model number and we use these models as a starting point for model selection for more parsimonious models of each type. As explained in la Cour et al (2019), 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). For a very brief description of the principles used in this gets model reduction strategy, please see la Cour et al (2019).

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 the 'ordinary' OLS standard errors as we find no strong indications of heteroscedasticity.

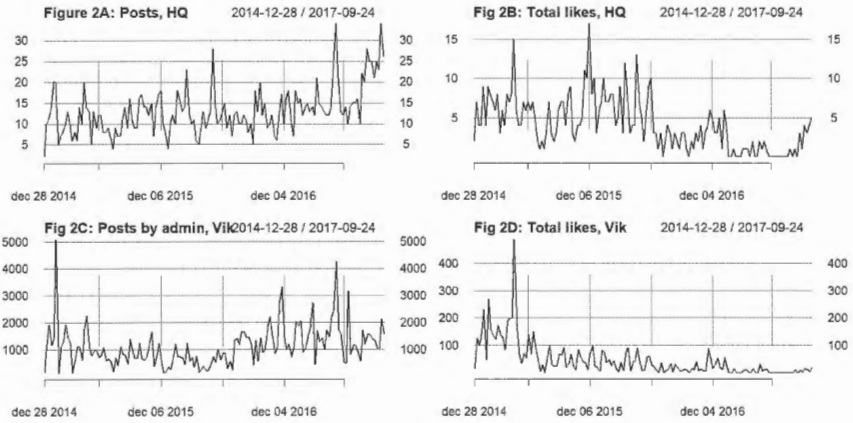
As we want to conduct an out-of-sample forecasting exercise based on a selected sub set of models we split our sample into a training or estimation sample and a test sample. The former runs from the beginning of 2015 and until the end of June 2017. The latter runs from the beginning of July until the end of September 2017.

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 for that).

Instead of visually trying to assess the presence of relationships between sales and the FB variables we will try to employ the automatic model selection routines explained above to backwards select from our (4) models.

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.



Finally, figures 3A to 3D show the most important remaining explanatory variables:

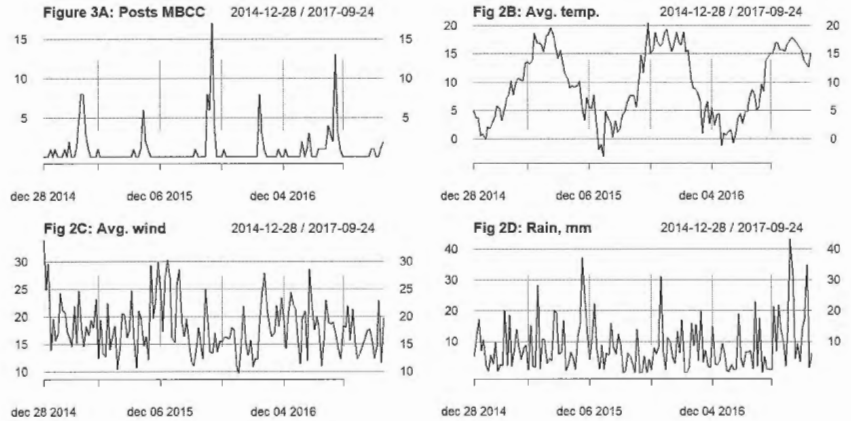


Figure 3A displays the development over time in the marketing effort related to the MBCC event. We see a pattern where kind of a first reminder is found the fall before the event and then more activity closer to the event itself. The remaining three graphs show the development in the weather variables: average temperture, average wind speed and the mm of rain. We expect that the weather variables may help capture the seasonality in the sales series.

Table 1: Descriptive summary statistics.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Sales	144	8,637.59	66.04	8,533.78	8,605.56	8,652.38	9,054.95
like_hq	144	1,122.97	746.65	125	637.8	1,436.2	5,080
like_vik	144	46.62	65.07	0	7.8	66.5	489
post_hq	144	13.13	5.64	2	10	15	34
post_vik	144	4.10	3.38	0	1	6.2	17
MBCCposts	144	0.93	2.36	0	0	1	17
temp_avg	144	9.56	6.15	-3	4.4	15.6	20
wind_avg	144	18.07	4.82	10	14.6	21	34
rain_mm	144	8.35	8.24	0.00	2.04	12.45	43.19

Note: 'Sales' are the anonymized, preprocessed numbers aggregated to the weekly frequency.

5. Estimation results of the regression models

Estimation results for the simplified regression models are shown in Table 2. We have chosen not to show the results of the full GUM specifications but focus instead just on the simplified or reduced versions to save space. The column numbers correspond to the equations (1) – (4) as mentioned earlier.

To save space we will not show the monthly and holidays dummies in the tables (even though some of them are picked by the gets procedure). A common thing is that the monthly dummies for April through September are positive and significant for models (1) and (2), i.e. the models where these monthly dummies are included. This is maybe not that surprising as people may tend to go more often for a beer in the bar in the warmer months than during the winter time. For models with holiday dummies included, the dummy for the Christmas week is significantly negative throughout. Other holiday dummies are now and then significant but not consistently so across the models. Notice that when included the MBCC dummies have always positive and very significant coefficients.

Turning to models (3) and (4) where the seasonal monthly dummies are replaced by the weather variables, we notice that the average temperature shows up in both cases and with a positive, very significant coefficient. Neither rain nor wind seem to matter in any of the models.

In Table 2 for model (4), we only see few significant FB variables. Surprisingly the selected FB variables show up with negative coefficients in the model and actually also the posts from the headquarters result in a negative long-run effect. When models like the ones discussed in this study is going to be used for predictive purposes, the R^2 of the models are also of interest. We notice that even with a very limited information set as for model (1) around 29% of the variation in sales are explained in-sample.

Adding the MBCC event dummies increases the R^2 substantially and already in model (2) it reaches around 90%. The explanatory power stays around this level both when the monthly seasonal dummies are replaced by the weather variables (eq. (3) and (4)) and also when the FB variables are added to the model (eq. (4)).

Table 2: Results from the final (specific) regressions for 'SALES'

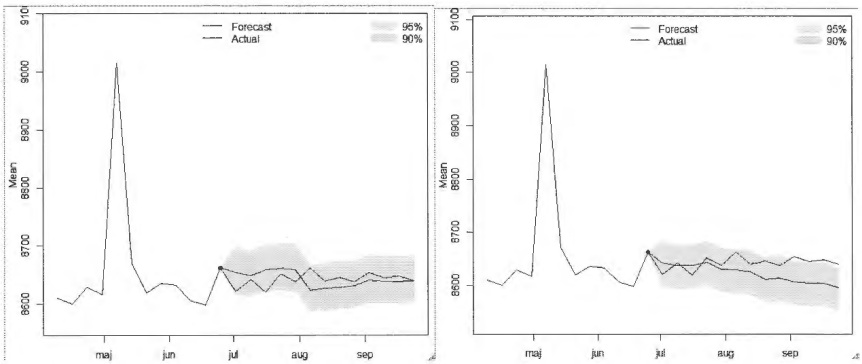
	(1)	(2)	(3)	(4)
ar2	-0.202**			
	(0.091)			
ar3	-0.150*			
	(0.088)			
ar1		0.186***	0.171***	0.220***
		(0.036)	(0.033)	(0.037)
ar4		0.141***		
		(0.041)		
Trend		-0.157***	-0.193***	
		(0.054)	(0.056)	
MBCC_15		336.023***	350.611***	352.524***
		(22.615)	(22.878)	(22.027)
MBCC_16		408.100***	409.306***	412.689***
		(22.641)	(22.759)	(22.001)
MBCC_17		391.521***	401.449***	411.199***
		(23.120)	(22.897)	(22.562)
post_hq1				-1.835***
				(0.473)
post_hq4				-1.135***
				(0.446)
temp_avg			2.472***	2.060***
			(0.358)	(0.359)
Constant	11,642.600***	5,809.510***	7,144.255***	6,498.320***
	(1,097.697)	(484.201)	(288.483)	(320.433)
Observations	127	127	127	127
LB(5), p-val:	0.597	0.794	0.069	0.885
ARCH(1), p-val:	0.662	0.239	0.196	0.115
R^2	0.327	0.915	0.900	0.908
Adjusted R^2	0.282	0.904	0.893	0.900

Note 1: *p<0.1; **p<0.05; ***p<0.01
Note 2: Standard errors in parentheses.
Note 3: A lot of dummies for months-of-year and also holidays are kept in some of the models but not displayed in Table 2. A constant term is present in all models.
Note 4: Standard F-tests for all models indicates strong significance.

6. An out-of-sample forecasting exercise.

In this section we will use model (2) and model (4) for a forecasting exercise. We keep the estimated coefficients from section 6 and calculate one-step-ahead forecasts for the weeks in the test sample (13 weeks in total). Notice that during the time period of the test sample there is no 'Mikkeller Beer Celebration Copenhagen' week so we can for this exercise rely on the models with the event dummies included. Also notice that all non-deterministic regressors except for the weather variables are lagged variables which makes the models suitable for forecasting in the present context. The weather variables are contemporaneous and therefore not directly suitable for a forecasting exercise. However for the present exercise we decided to rely on the actual values keeping in mind that the weather forecast for the coming week could have replaced these values in future research. As we do not have access to historical weather forecasts of this type and as it seems that these variables more or less substitute for seasonality (= e.g. our monthly dummies), we are not too worried about this shortcoming.

Figure 4: Forecasts and actual values for model (2) and model (4).



Model (2) Model (4)

As seen in Figure 4, left part, the actual values lie within the 90% confidence bounds of the forecasts for model (2) most of the time. However the actual temporal pattern of the forecast line is much smoother than for the actual values. Overall the forecasts are not too bad for a model based on a very limited and rather deterministic set of regressors.

In the case of model (4), right part of Figure 4, there seem to be some kind of bias in the forecasts as the actual values towards the end of the test sample seem to move outside the upper 90% confidence bounds. Hence replacing the monthly dummies with the weather information (average temperature) and also adding selected FB variables does not seem to improve the out-of-sample predictive power of the model.

Based on figure 4 we conclude that model (2) seems to be superior when compare to . model (4) in the out-of-sample forecasting exercise. This conclusion is supported by a selection of numerical measures of forecasting accuracy, see table3.

Table 3. Numerical measures of out-of-sample forecast accuracy

Accuracy measures	Model 2	Model 4
RMSE	20.70	30.26
MAE	16.46	26.39
MAPE	0.0002	0.0031

Note: RMSE is the root mean square error, MAE is the mean absolute error and MAPE is the mean absolute percentage error.

7. 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 insignificant 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:

$$S_t = -(S_{t-1} + \ldots + S_{t-6}) + \zeta_t$$

The yearly variation is modelled by cycle components; in this case by cosines for approximately a whole calendar year, a half year and a third of a year.

In total these component lead to the basic model:

$$y_t = \mu_t + S_t + \text{cycles} + \varepsilon_t$$

All remainder terms, ε_t, η_t, and ζ_t, are assumed to be mutually independent white noise series. But in the initial model applications such assumptions are far from valid. The component 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; in this case a model with fixed seasonal dummies is found if var(ζ_t) = 0. In this application the variance for the level component error, var(η_t) is also small and insignificant, but in the initial model this term is however included.

In the first round the parameters of this model are estimated by the Kalman filter extended by a further smoothing estimation algorithm. The smoothing is applied in order to estimate the true level of all components for each observation using both past and future observed values in the calculation.

In this situation the result is a smoothed value defined as the estimated value of

$z_t = \mu_t + S_t + \text{cycles}.$

Note that the series, z_t, is a smoothed version of the original series, y_t, without irregularities and outliers. The residuals are then defined by

$r_t = y_t - z_t$

If the residual r_t is numerically large (larger than three items the standard deviation) the observed value, y_t, is set as missing. This gives a new series $y^1{}_t$ in which many irregularities from the original series are removed.

This smoothing process could be applied once again as the Kalman filter and the smoothing algorithm for estimation of the components work even for time series with missing observations. The result is a more smoothed version, $z^2{}_t$ of the series z_t and a new series, $y^2{}_t$, with more missing values than the series $y^1{}_t$ .This procedure is applied two times more leading a smooth time series , $r^*{}_t$ and new series $y^*{}_t$ where 43 observations where outliers are replaced by a missing value.

The irregularity residuals

$r^*{}_t = y^*{}_t - z^*{}_t$

are defined is order to see the difference between the actual observation and a very smoothed value of the series.

8. Analyzing the irregularity residuals for daily data

In this section daily data is used. The series which is transformed linearly for confidence reasons and a log-transformation is applied. Figure 5 presents the original series, y_t, and the smoothed series $y^*{}_t$, while the next Figure 6 gives a plot of the irregularity series $z^*{}_t$.

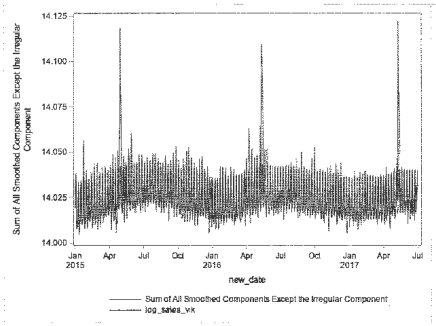


Figure 5

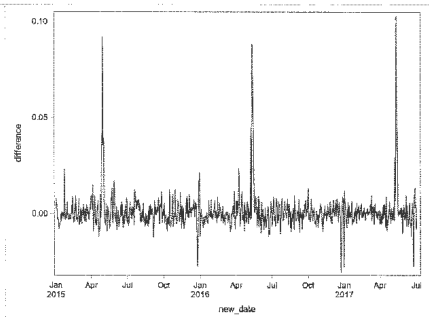


Figure 6

The idea is to use the total battery of more than 30 potential explanatory variables in order to explain as much as possible of the irregular residuals as possible.

From the plot of the series of irregular residuals it is clear that the few negative residuals with large absolute values are for observations late December and early January while the largest residuals are for May or early June. This is expected and easily explained as a result of celebrations around Christmas and New Year and the beer festival. These effects are in a model explained by dummy variables as such effects are well known – and they are of course also to be expected in the years to come. One strategy is to dummy these findings out and see how to explain the remaining residuals.

An ordinary regression model is then fitted by Ordinary Least Squares to explain these irregular residuals by dummies for the holiday season and the beer festival. Moreover some other well established reasons for large or low sales of beer like some further Danish holidays, the first of May etc. are included among the extra 24 explanatory variables. The residuals for this regression is plotted in the next Figure 7.

In this plots all well-known reason for important irregular residuals are counted for leaving space to see whether weather data and social media data can be used to explain special events. The standard deviation and also the numerical values of outliers are of course much smaller than for the original irregular residuals defined later.

This series of residuals is then fitted by weather data and social media data. The results are however disappointing as no important significance is found. The overall R^2 is around 0.02 if all these weather and social media variables are used in a large model. In a more detailed analysis in order to find a reasonable small number of exogenous variables it could be that the total number of posts by the specific bar in Viktoriagade, $p = 0.084$, and the total quantity of rain measured in mm, $p = 0.049$ seems to have a significant parameter; however the R^2 for such models are however around 0.01.

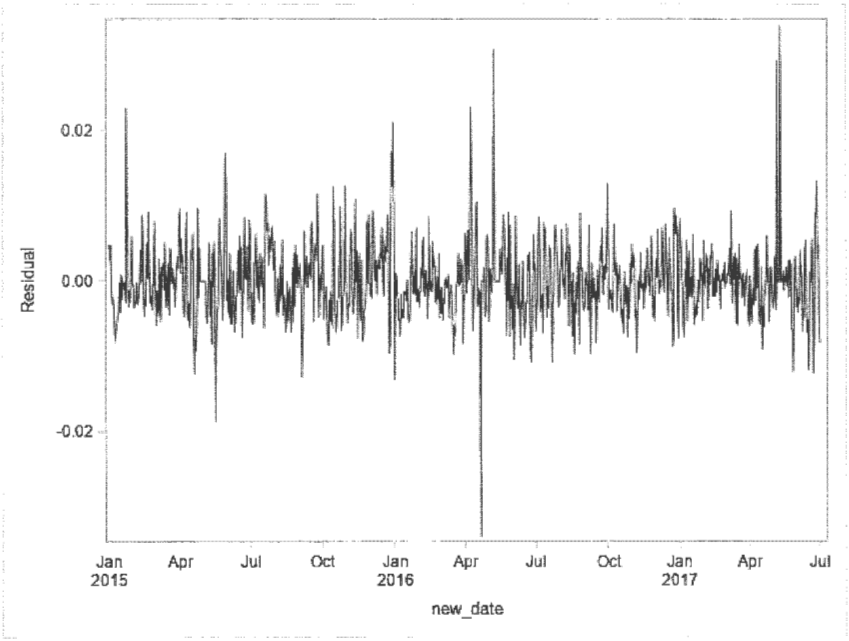


Figure 7

In this approach all external effects of the calendar and the beer festival are removed prior to the analysis. In an analysis where such effects are not included and only the weather and the social media variables are applied R^2 is 0.07 and the number of posts from the company and the number of likes at the specific bar are highly significant. That tells perhaps just that the number of posts by the Company are large at certain days; probably in the days of the beer festival. In a model where all available variables are applied simultaneous, R^2 is as large as $R^2 = 0.71$ as the calendar effects and the beer festival explains a lot, but the parameters to the social media data and the weather variables are all insignificant.

In this dataset it is very natural that the marketing efforts are extraordinary high at days with expected large sales like evenings before non-working days and of course also during the beer festival. In statistical terms this leads to multi collinearity. But the conclusion from this analysis is that the social media and the weather have only very small impact on the sales if any impact at all.

10. Models for the right tail of the distribution

A series problem in the analyses in paper is the non-normality in all regressions and time series analyses in this paper. All error distributions have heavy right tails, while the left tails are much smaller. The logarithmic transformation which is applied in the

unobserved component models is far from enough to remedy this, see Figure 6 of the irregularity residuals and the Figure 7 for model residuals. Moreover all management efforts try to increase the sales and positive outliers are in fact the goal!

In this section the distribution of the right tail will be studied in much more detail by PROC SEVERITY in SAS.

First three possible distributions are tried for the right hand tail of the distribution of the right tail: Weibull, Pareto and Burr. Only the right tail is considered as the distribution is censored at the 75 quantile. This means that only the 25% largest irregular residual, r^*_t, are used in the analyses. The model are compared using information criteria like Akaike Information criterion and Swartz Bayesian Information Criterion. The conclusion is clear, the Burr distribution gives the best fit. As many Burr distributions exist, this is the Type XII Burr which is also called the Singh-Maddala distribution or the generalized log-logistic distribution. This distribution has the density

Name	Distribution	Parameters	PDF (f) and CDF (F)
BURR	Burr (Type XII)	$\theta > 0$, $\alpha > 0$, $\gamma > 0$	$f(x) = \frac{\alpha\gamma z^{\gamma}}{x(1+z^{\gamma})^{(\alpha+1)}}$ $F(x) = 1 - \left(\frac{1}{1+z^{\gamma}}\right)^{\alpha}$

Moreover a scale parameter, θ, could be included. This scale parameter is allowed to be a linear function of exogenous variables. The table give the most important information criteria for combinations of exogenous variables. By trial and error it is concluded that only the number of likes of the specific bar adds more information to the model bases on all calendar effects and the effect of the festival. This extra effect is significant when judged by the likelihood and by uncorrected AIC, but less important judged from the corrected AIC and Swartz Bayesian Information Criterion.

Model	Distribution	-2LogLikelihood	AIC	AICC	BIC
No exogenous	Burr	-1307	-1301	-1301	-1289
Only calendar and festival	Burr	-1479	-1401	-1393	-1242
Only likes for the specific bar	Burr	-1485	-1405	-1396	-1242
All possible	Burr	-1492	-1396	-1384	-1201

The parameters are estimated as; including only the important regression parameter

Parameter	DF	Estimate	Standard Error	tValue	Pr > |t|
Theta	1	0.00676	0.00281	2.41	0.0164
Alpha	1	2.83222	1.36773	2.07	0.0390
Gamma	1	1.68592	0.22359	7.54	<.0001
Likes for the specific bar	1	0.00567	0.00251	2.26	0.0246

The distribution is heavy tailed as only moments up to order $\alpha\gamma = 4.87$ exist. The censored Burr distribution is plotted in Figure 8. The scale parameter, θ is defined as a linear expression

$\theta = 0.00676 +$ many regression times including $0.00567 \times$ Likes for the specific bar

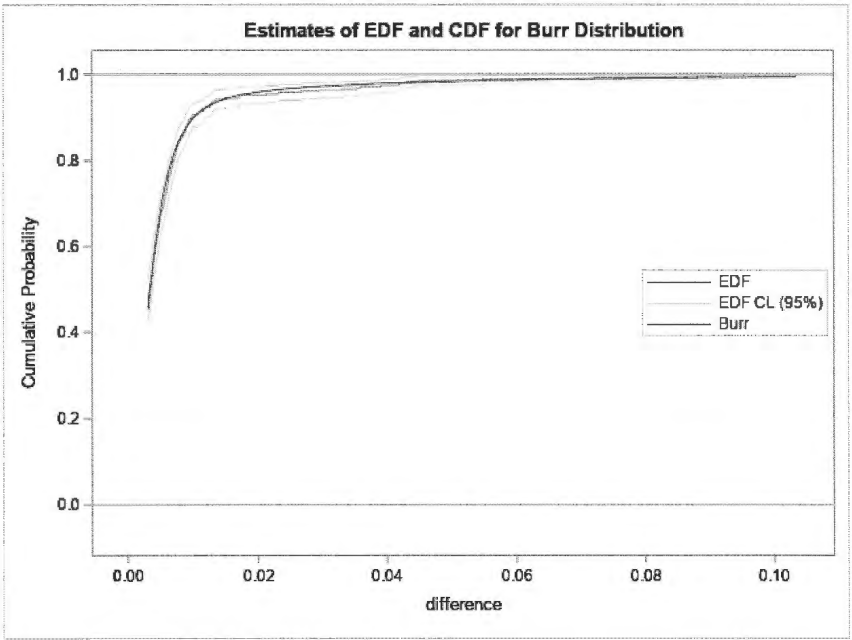


Figure 8

11. Summary and conclusion

In this paper we have pursued our previous efforts to determine a forecasting model for sales of the Danish microbrewery Mikkeller; this year even by inclusion of weather data. We tried a strategy with lagged sales to model the autocorrelation of the in the sales series and a suite of variables for deterministic outside factors like Facebook variables and weather variables. These exogenous variables were also applied to a smoothed version of the sales series. Moreover they applied to a model for the right tail of the sales distribution in a setup of a heavy tailed Burr distribution

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 Facebook variables and weather variables generally did not improve much to R^2 or information criteria compared to the benchmark models.

12. 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

La Cour, L., Milhøj, A. , Vatrapu, R. and Buus Lassen, N. (2019). 'Mikkeller – a Third Attempt', in P. Linde (Ed.) *Symposium i Anvendt Statistik*, Page 126-140

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

Survival after cancer for twins aged 70+

Martin D. Villumsen[1], Kaare Christensen[1,2], Marianne Ewertz[3], and Jacob Hjelmborg[1,2]

[1]*Department of Epidemiology and Biostatistics, University of Southern Denmark, Denmark*
[2]*The Danish Twin Registry, University of Southern Denmark, Denmark*
[3]*Oncology Unit, Department of Clinical Research, Odense University Hospital, Odense, Denmark*

Aim

The aim of this study is to explore the impact of cancer on mortality risk for the elderly by a population-based study on twins. For Danish born twin pairs with both twins alive at first cancer diagnosis in the pair, firstly, mortality for the twins with cancer first in pair is compared to that of the co-twins and, secondly, excess mortality is estimated in matched design. Cancer is defined as any cancer except non-melanoma skin cancer.

Danish Twins

Danish born, complete twin pairs recorded in the Danish Twin Register at end–2009 and with one or more cancer diagnoses listed in Danish Cancer Register from 1943 to end–2011 are included in this study if both members were alive at age six and at the date of the first cancer diagnosis in the pair. In total, 7990 pairs where 1768, 5710, and 512 pairs are monozygotic, dizygotic respectively of unknown zygosity.

The population-based Danish Twin Register dates from 1954 and covers twin and multiple births from 1870 and forth (Skytthe, et al. 2011). The identification of twin individuals is carried via church records for early cohorts, the Danish Civil Registration System for births after 1st April 1968, and the Medical Birth Registry from 1973 (Skytthe, et al. 2006). Zygosity of same-sex twin pairs is categorised using four standard questions, an approach with a misclassification rate under five percent (Christiansen et al., 2003).

Mortality and cancer incidence for Nordic twins and the background population are comparable, and mortality after the age of six of Danish twin individuals is akin that of the background population; infant mortality among twin individuals is higher when compared to the background population (Skytthe et al., 2019) (Christiansen et al., 1995) (Kleinman et al. 1991). For twins themselves, no substantial or systematic difference in survival after infancy has been found between monozygotic and dizygotic twin individuals (Hjelmborg et al., 2019). Therefore, as a twin individual shares age, genes, and

socio-economic factors with its co-twin, it seems much in evidence to describe the impact of a cancer diagnosis via a register study on twins.

Cancer diagnoses in twins

About 12% of the Danish born twin individuals are diagnosed with cancer between the beginning of 1943 until the end of 2011. Fewer diagnoses are found in the 1940s and 1950s when compared to later decades, a distinctive leap is observed after the initialisation of the Civil Registration System in 1968, and, generally, the cancer diagnoses frequency increase with age. The Lexis diagram in Figure 1 illustrates first cancer diagnoses for Danish born twins over time.

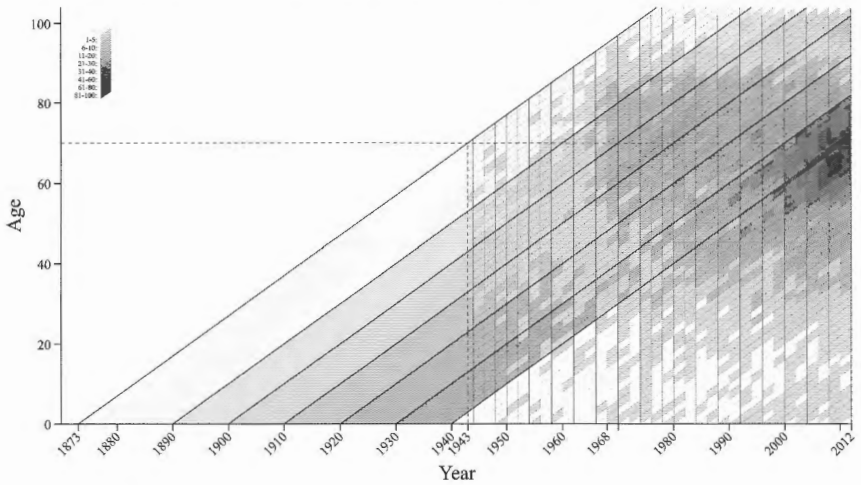


Figure 1: The Lexis diagram illustrates the number of first cancer diagnoses in twin individuals from birth cohorts 1873-2011. The number of twin individuals registered with cancer in the Danish Cancer Register (initiated in 1943) increases with age.

It became possible after the introduction of the Danish Civil Registration System, with very few exceptions, to identify all Danish citizens alive at the beginning of the initialization and onwards. Therefore, 1st April 1968 is a date whereafter one expects an increase in recorded cancer incidences for twins; an increment at 1968 is highly apparent with the Lexis diagram in Figure 1.

Old Cohorts

Old birth cohorts allow for high percentage of full follow-up. Table 1 shows average remaining life difference in years between twins with a cancer diagnosis first in the pair and their co-twins. The pairs

are taken from birth cohorts with almost full follow-up; with both twins alive at first cancer diagnosis; one or more cancer diagnoses in the pair; and first cancer diagnosis at age 70 or later.

Born	FFU	Mean	95% CI
1873–1890	49 (100%)	4.38 years	(2.87 5.90)
1891–1900	225 (100%)	5.80 years	(4.94 6.67)
1901–1910	329 (99.7%)	6.18 years	(5.40 6.97)
1911–1920	333 (98.1%)	5.87 years	(5.07 6.68)

Table 1: For birth cohorts with nearly full follow-up (FFU) and in pairs with 1+ cancer, mean survival difference between co-twin (not cancer first) and twin (cancer first) from time of first cancer.

Survival Curves for 70+

To examine, in more detail, differences in life expectancies across decades for the elderly, the following four births-cohorts are now considered, 1898*–1910, 1911–1920, 1921–1930, and 1931–1941 (*: from 2nd April). With this setup, cancer ranges from 2nd April 1968 to 31st December 2011 as twin individuals are included here if time of first cancer in the pair was at age 70 or later.

	All	1898–1910	1911–1920	1921–1930	1931–1941
Total, n (%)	3218	812 (25%)	728 (23%)	790 (25%)	888 (28%)
Age at FU, median (IQR)	79.7 (75.5 85.1)	83.2 (77.5 88.5)	83.1 (77.5 89.3)	82.2 (77.4 87.9)	75.7 (73.3 78.4)
Dead during FU, n (%)	2 362 (73.4)	810 (99.8)	671 (92.2)	499 (63.2)	255 (28.7)
Age at FCIP, median (IQR)	74.6 (72.1 78.1)	75.8 (72.6 80.3)	75.9 (73.0 79.4)	75.5 (72.7 79.1)	72.5 (71.2 74.9)
Cancer, n (%)					
No	1365 (42.4)	331 (40.8)	298 (40.9)	329 (41.6)	407 (45.8)
Yes, but not CF	244 (7.6)	75 (9.2)	66 (9.1)	66 (8.4)	37 (4.2)
CF	1609 (50.0)	406 (50.0)	364 (50.0)	395 (50.0)	444 (50.0)
Sex, n (%)					
Male	1552 (48.2%)	374 (46.1)	324 (44.5)	372 (47.1)	482 (54.3)
Female	1666 (51.8%)	438 (53.9)	404 (55.5)	418 (52.9)	406 (45.7)
Zygosity, n (%)					
MZ	822 (25.5)	190 (23.4)	224 (30.8)	258 (32.7)	150 (16.9)
DZ	2250 (69.9)	584 (71.9)	474 (65.1)	494 (62.5)	698 (78.6)
UZ	146 (4.5)	38 (4.7)	30 (4.1)	38 (4.8)	40 (4.5)

Table 2: Population characteristics of Danish born twin individuals with one or more cancer diagnoses in the pair at age 70 or later, both members alive at the first diagnosis, and born from April 2nd 1898 to December 31st 1941.

Censored observations frequently occur when more recent birth cohorts are considered. Kaplan Meier survival curves are shown in Figure 2A for the four birth cohorts; in total, 3218 twin individuals from complete twin pairs are included with characteristics listed with Table 2. The survival curves show

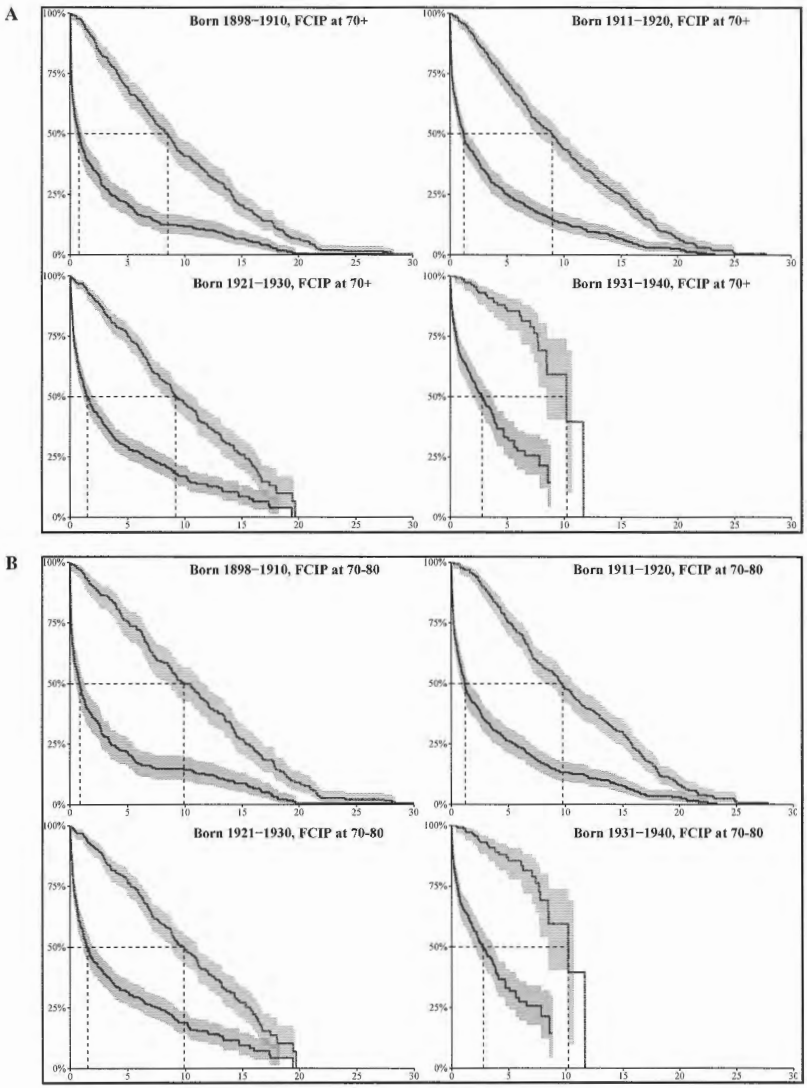


Figure 2: Kaplan-Meier survival estimates for twins from twin pairs with first cancer in pair (FCIP) at age 70 or more (**A**) and at age 70–80 (**B**). Lower and upper curves are for twins with cancer first in pair respectively not first in pair. Time in years after first cancer in pair is used as time scale.

improvement in median survival for both cancer patients and the co-twins. However, median life expectancies are not comparable across the four cohorts due to the end of follow-up at 31st December 2011. For example, in the birth-cohort 1931–1941 age 81 is an upper limit for detecting pairs with first cancer whereas in the three earlier cohorts, pairs with first cancer at higher ages can be detected. Since the life expectancy is highly affected by the age at diagnosis, the change in life expectancy across the four cohorts may be confounded due to selection bias from age at diagnosis. For comparison, Figure 2B shows Kaplan Meier survival curves for the same four cohorts but for pairs with cancer diagnosed restricted to the age span 70–80, and thus survival curves (upper) for the co-twin become somewhat lifted for the three early cohorts.

There are indications of non-proportional hazards with the curves shown in Figure 2. The possible time-varying effect on getting cancer first is assessed by Aalen's additive model from which the mortality risk seems substantial and immediately but over time strongly decreasing. Subsequently splitting of the time scale for estimating separate hazard ratios on different time intervals in Cox proportional hazard model leads one to introduce the function

$$t \mapsto \exp\left(-\sqrt{t}\right) \tag{1}$$

to encapsulate the time-varying effect.

Table 3 summarize the effect of getting cancer first (CF) for males and females at one and five years after the first diagnosis in pair for elderly twins. The estimates are from a Cox proportional hazard model adjusting for sex, zygosity, age at first diagnosis in pair (70–75, 75–80, or 80+), birth cohort, and interacting effect of time on sex and CF as described by (1). Though the mortality risk is lower for females when compared to males (HR (95% CI) 0.63 (0.55 0.72)), there is a higher risk, which decreases over time, for females than for males when comparing individuals with CF to those not with CF (see Table 3). Figure 3A and 3B illustrate the differences for male and females via predicted survival curves for monozygotic individuals from birth cohorts 1895–1910 respectively 1921–1930.

Birth Cohort	AFD: 70–75	AFD: 75–80	AFD: 80+
1 year after first cancer in pair			
Male			
1898–1910	8.50 (6.74 10.7)	6.15 (4.89 7.74)	4.40 (3.50 5.52)
1911–1920	8.28 (6.52 10.5)	5.99 (4.74 7.57)	4.28 (3.30 5.55)
1921–1930	7.87 (6.20 9.97)	5.69 (4.47 7.23)	4.07 (3.11 5.32)
1931–1940	8.78 (6.24 12.3)	6.35 (4.38 9.02)	4.54 (3.06 6.74)
Female			
1898–1910	10.6 (8.20 13.8)	7.70 (5.97 9.92)	5.50 (4.27 7.08)
1911–1920	10.4 (7.93 13.5)	7.50 (5.79 9.70)	5.35 (4.04 7.10)
1921–1930	9.84 (7.55 12.8)	7.12 (5.47 9.27)	5.09 (3.81 6.80)
1931–1940	11.0 (7.64 15.8)	7.94 (5.39 11.7)	5.68 (3.76 8.57)
5 year after first cancer in pair			
Male			
1898–1910	2.07 (1.73 2.49)	1.50 (1.24 1.82)	1.07 (0.88 1.32)
1911–1920	2.02 (1.66 2.45)	1.46 (1.19 1.79)	1.04 (0.82 1.33)
1921–1930	1.92 (1.56 2.36)	1.39 (1.11 1.74)	0.99 (0.76 1.29)
1931–1940	2.14 (1.51 3.03)	1.55 (1.06 2.28)	1.11 (0.73 1.67)
Female			
1898–1910	2.21 (1.84 2.67)	1.60 (1.32 1.95)	1.14 (0.93 1.41)
1911–1920	2.16 (1.77 2.63)	1.56 (1.27 1.91)	1.11 (0.87 1.42)
1921–1930	2.05 (1.66 2.53)	1.48 (1.18 1.86)	1.06 (0.81 1.38)
1931–1940	2.29 (1.61 3.24)	1.65 (1.13 2.43)	1.18 (0.78 1.79)

Table 3: Hazard ratio (95% CI) of CF vs. not-CF at 1 and 5 year after first cancer in pair for males and females. AFD: age at first cancer diagnosis in pair. Cluster robust standard errors are used to account for dependencies within twin pairs.

Excess risk by matched design

With a different setting, 5680 Danish born same-sex complete twin pairs with one or more cancer diagnoses in the pair and both alive at first cancer after the age of six are now considered. This set of twin pairs forms a cohort with a natural matched structure on age, sex, and socio-economic factors between the twins with cancer first in pair and their co-twins. Excess mortality risk of cancer is estimated by a proportional excess risk model where the hazard for the twins with cancer first is represented as a sum of a population baseline hazard and a proportional excess hazard (Boschini, et al. 2019). The approach assumes that the risk for cancer survivor is the sum of a cluster-specific background risk defined on the age time scale and an excess term defined on the time since exposure time scale.

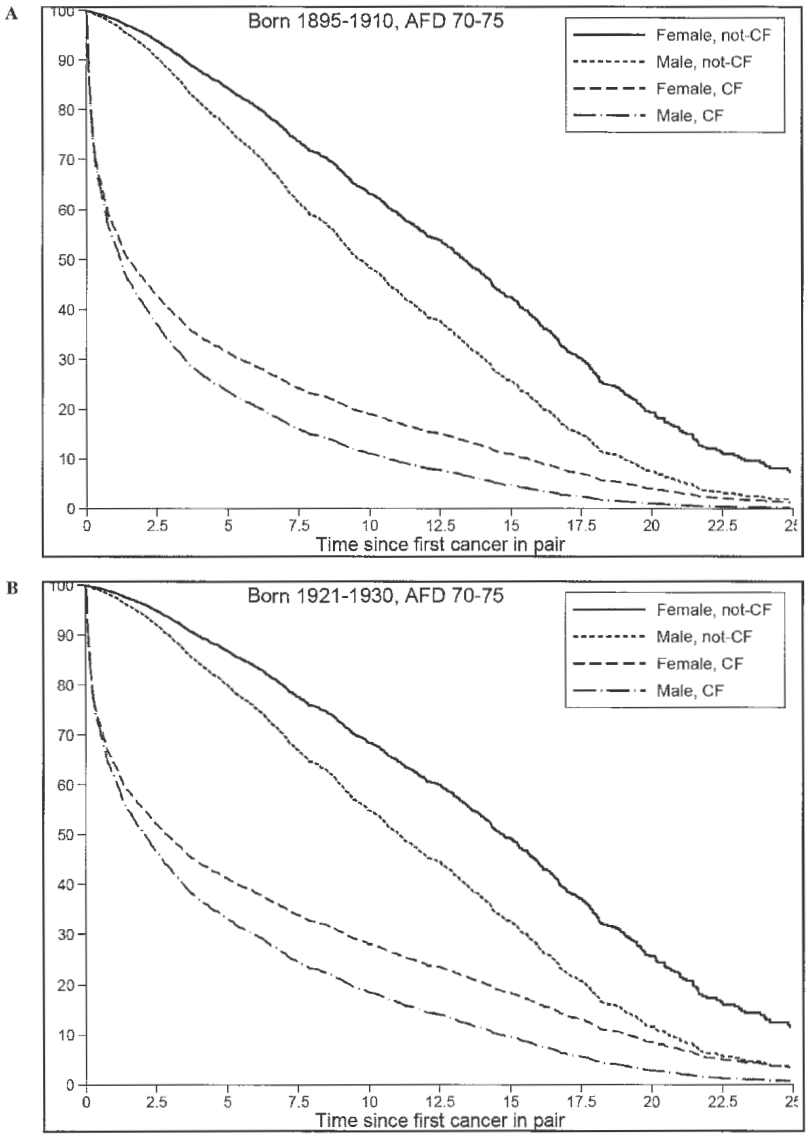


Figure 3: Predicted survival curves for male and female Danish monozygotic twin-individuals from time of first cancer-diagnosis in the pair; both members alive at the date of diagnosis and age at first diagnosis in the pair (AFD) ranges in 70–75. Individuals with cancer first (CF)

	All
Total, n (%)	11 360
Year of first diagnosis in pair	
1943–1st April 1968	590 (5.2)
2nd April 1968–1980	2 070 (18.2)
1981–1990	2076 (18.3)
1991–2000	2758 (24.3)
2001–2011	3866 (34.0)
Age at first diagnosis in pair, n (%)	
6–39	1174 (10.3)
40–69	7298 (64.2)
70+	2888 (25.4)
Sex, n (%)	
Male	5136 (45.2)
Female	6224 (54.8)
Zygosity	
Monozygotic	3528 (31.1)
Same-sex dizygotic	6 842 (60.2)
Unknow	990 (8.7)

Table 4: Population characteristics of Danish born twins from same-sex pairs with one or more cancer diagnoses in the pair at age six or later and both members alive at the first diagnosis.

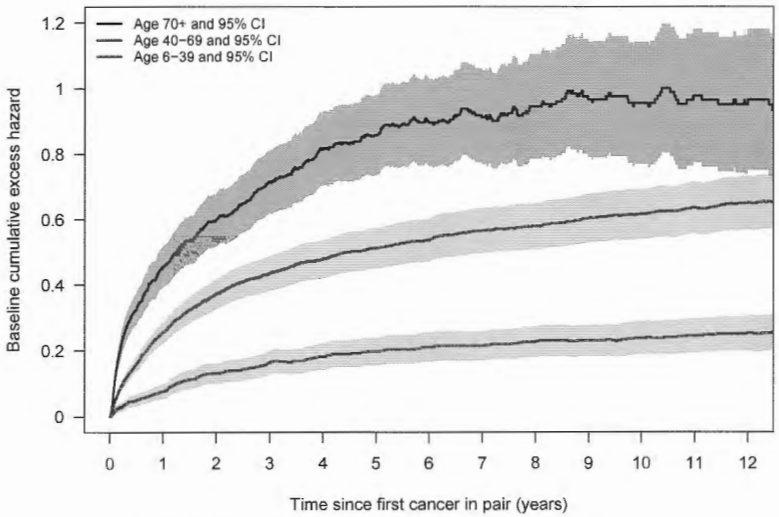


Figure 4: Excess cumulative mortality hazards stratified by age for monozygotic male with first cancer diagnosis in 2001–2011.

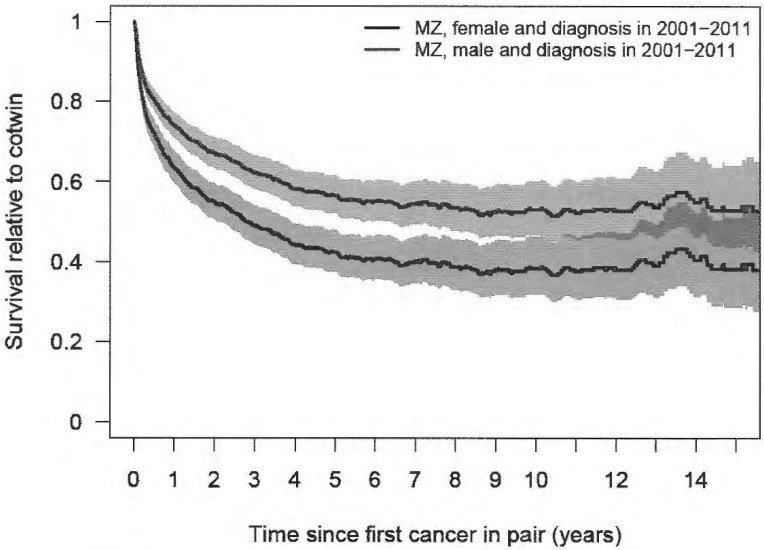


Figure 5: Survival curve for twin with CF relative to co-twin for elderly (70+) monozygotic (MZ) male and fermale, with year of first diagnosis in 2001–2011.

Adjusting for sex, zygosity, and year of first diagnosis in pair, estimates of excess cumulative mortality hazards are obtained by stratifying on age intervals 6–39, 40–69, and 70+ with respect to first diagnosis in pair; figure 4 shows excess cumulative hazards for baseline (monozygotic, male, 2001–2011 year of first diagnosis in pair). On survival scale, Figure 5 shows relative (to co-twin) survival curves for elderly monozygotic male and female with cancer first in pair in 2001–2011. A state of equilibrium seems to be reached after five to seven years.

Short summary

The impact of cancer, which has here been defined as any cancer diagnosis except non-melanoma skin cancer, is explored via a population-based study on twins where (ideal) matching between twin and co-twin opens up for handling two time scales, age and time since diagnosis, when estimaing excess mortality risk. The seriousness of cancer declines somewhat over birth cohorts, yet the immediate hazard after a cancer is high; one year after a cancer at age 70, the increased hazard is about tenfold, at age 80 it is fivefold. The relative risk stabilises for the 70+ after about five to seven years.

References

Boschini, C., K.K. Andersen, and T.H. Scheike, *Excess risk estimation for matched cohort survival data.* Stat Methods Med Res, 2019. 28(10-11): p. 3451-3465.

Christensen, K., et al., *Mortality among twins after age 6: fetal origins hypothesis versus twin method.* BMJ, 1995. 310(6977): p. 432-6.

Christiansen, L., et al., *Age- and sex-differences in the validity of questionnaire-based zygosity in twins.* Twin Res, 2003. 6(4): p. 275-8.

Kleinman, J.C., M.G. Fowler, and S.S. Kessel, *Comparison of infant mortality among twins and singletons: United States 1960 and 1983.* Am J Epidemiol, 1991. 133(2): p. 133-43.

Hjelmborg, J., et al., *Lifespans of Twins: Does Zygosity Matter?* Genes (Basel), 2019. 10(2).

Skytthe, A., et al., *The Danish Twin Registry in the new millennium.* Twin Res Hum Genet, 2006. 9(6): p. 763-71.

Skytthe, A., et al., /The Danish Twin Registry.(Scand J Public Health, 2011. 39(7 Suppl): p. 75-8.

Skytthe, A., et al., *Cancer Incidence and Mortality in 260,000 Nordic Twins With 30,000 Prospective Cancers.* Twin Res Hum Genet, 2019. 22(2): p. 99-107.

Medianens varians og efficiens – et simulationsstudie.

En afsøgning af den mest efficiente estimator af medianen for normalfordelte observationsværdier.

Steen Andersen

Studielektor, Institut for Økonomi, BSS, Aarhus Universitet, e-mail: sta@econ.au.dk

***Abstract**: Medianestimatet har sin relevans eller berettigelse ved asymmetriske fordelinger som f.eks. indkomst samt i situationer, hvor rangordningen er enkel, og hvor målingen er kompleks. Dette studie tager udgangspunkt i en situation, hvor man meget enkelt og entydigt kan rangordne, men hvor en måling er meget omkostningsfuld. Det antages som udgangspunkt, at observationsværdierne er normalfordelte.*

De væsentligste fund i dette studie er, for det første, at den varians af medianen, som mange lærebøger i statistik anfører; π gange σ^2 divideret med 2 gange stikprøvestørrelsen - er upward biased, specielt ved små stikprøvestørrelser. Et andet fund er, at et gennemsnit af de to midterste observationsværdier ikke er den mest efficiente estimator af medianen. Dette leder hen til det tredje fund; at de to observationsværdier, - ud af n observationsværdier, der kan bidrage med det mest efficiente gennemsnit, - er nedre og øvre kvartil. Variansen af denne estimator kan vises at være cirka π gange σ^2 divideret med 2,5 gange stikprøvestørrelsen. Inddrager man også den midterste kvartil ved gennemsnitsberegningen så har denne estimator en varians på cirka π gange σ^2 divideret med 2,67 gange stikprøvestørrelsen.

Afslutningsvis anføres et par enkelte overvejelser om konsekvenserne af diverse afvigelser fra normalitetsantagelsen.

Indledende overvejelser og vurderinger.

Medianestimatet er defineret som den midterste observationsværdi i et rangordnet datasæt når man har et ulige antal observationer og er defineret som et gennemsnit af de to midterste observationsværdier når man har et lige antal observationer.

Variansen af medianen afhænger af populationens fordeling og af antal observationer (n). I Maritz og Jerret (1978) og i Arnold et. al. (2008) er mere generelle udtryk for variansen af medianen beskrevet.

I mange lærebøger er det anført, at variansen af medianen er $V(Md)=\frac{\pi\cdot\sigma^2}{2\cdot n}$ når populationen er normalfordelt, $X \square N(\mu,\sigma^2)$. Men ved n=1 er $V(Md)=V(X)=\sigma^2$ og ved n=2 er $V(Md)=V(\bar{X})=\frac{\sigma^2}{2}$, hvilket betyder, at en korrektion for blandt andet n bør inddrages.

Et bud på en korrektion, fundet ved simulationer samt ved trial and error, kunne være:

$$V(Md_{Kor})=\frac{\pi_n\cdot\sigma^2}{2\cdot n\cdot\left(\frac{n^2+\frac{n}{3}}{n^2-\frac{n}{3}}\right)}=\frac{(3\cdot n-1)\cdot\pi_n\cdot\sigma^2}{2\cdot n\cdot(3\cdot n+1)} \text{ for ulige n}$$

$$V(Md_{Kor})=\frac{\pi_n\cdot\sigma^2}{2\cdot n\cdot\left(\frac{n^2+\frac{n-2}{2}}{n^2-\frac{n}{2}}\right)}=\frac{(2\cdot n-1)\cdot\pi_n\cdot\sigma^2}{2\cdot(2\cdot n^2+n-2)} \text{ for lige n}$$

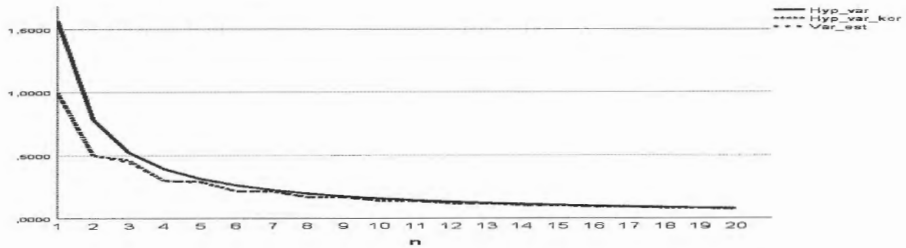
hvor $\pi_n=4\cdot\sum_{i=1}^{n}(-1)^{i+1}\cdot\frac{1}{2\cdot i-1}$ $\qquad \pi=4\cdot\sum_{i=1}^{\infty}(-1)^{i+1}\cdot\frac{1}{2\cdot i-1}$

$$\pi_3=4\cdot\sum_{i=1}^{3}(-1)^{i+1}\cdot\frac{1}{2\cdot i-1}=4\cdot\left(\frac{1}{1}-\frac{1}{3}+\frac{1}{5}\right)=4\cdot\frac{15-5+3}{15}=\frac{52}{15}=3,47$$

Nedenfor er de estimerede varianser (Var_est) af medianen, baseret på 1 million simulationer med stikprøvestørrelser fra 1 til 20 og $X \sim N(\mu = 0, \sigma^2 = 1)$, illustreret.

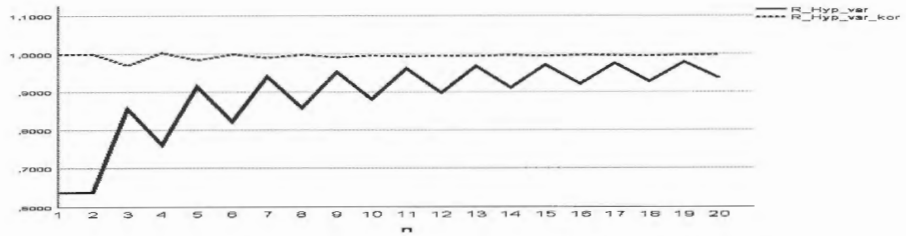
Stort set sammenfaldende med disse varianser ses de hypotetiske varianser (Hyp_var_kor) for medianen, hvor korrektionen er inddraget.

Den traditionelle hypotetiske varians (Hyp_var) for medianen er desuden illustreret for n=1 til n=20.



Den traditionelle hypotetiske varians $V(Md) = \frac{\pi \cdot \sigma^2}{2 \cdot n}$, vurderes på grundlag heraf til at være upward biased, hvorimod den korrigere variansestimator $V(Md_{Kor})$ vurderes til at være unbiased.

Dette forhold fremgår tydeligere når de estimere varianser divideres med de hypotetiske varianser:



Bortset fra en afvigende tendens ved n=3 og n=5 tyder $V(Md_{Kor})$ på at være unbiased.

Såfremt man antager, at π_n korrektionen er korrekt og dernæst regner baglæns for at fastslå hvad nævner-korrektion skulle være får man følgende baseret på 1 million simulationer:

n	$\frac{\pi_n \cdot \sigma^2}{2 \cdot n}$	s^2_{Md}	$\frac{\pi_n \cdot \sigma^2}{2 \cdot n} \Big/ s^2_{Md}$	Korrektion $\left(\frac{n^2+\frac{n}{3}}{n^2-\frac{n}{3}}\right)$ for ulige n $\left(\frac{n^2+\frac{n-2}{2}}{n^2-\frac{n}{2}}\right)$ for lige n
1	2,00000	0,99888	2,00224	2
2	0,66667	0,49956	1,33450	1,333
3	0,57778	0,44828	1,29889	1,250
4	0,36190	0,29806	1,21419	1,214
5	0,33397	0,28670	1,16488	1,142
6	0,24800	0,21450	1,15620	1,152

Ud fra ovenstående vurderes det, at der ikke umiddelbart er andre "heltalskorrektion", der ligger lige for. Korrektionerne for n=3 og 5 er de som afviger mest, - en alternativ korrektion for n ulige kunne være: $\left(\frac{n^2+\frac{n-1}{3}}{n^2-\frac{n}{2}}\right)$.

Der vil fortsat være større afvigelser end ved korrektionerne for n lige.

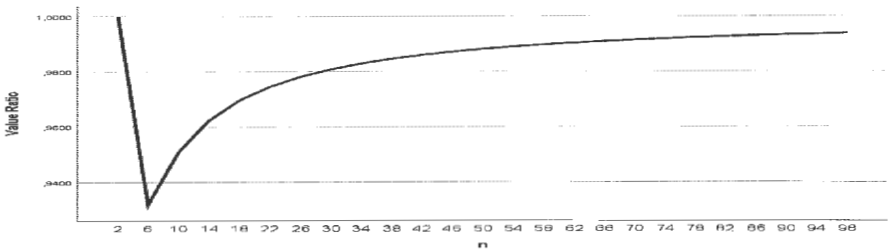
Efficiensvurderinger ved middelværdiestimatorer.

Antag, at man står i en situation, hvor X er normalfordelt og hvor n/2 er et ulige tal.

Antag, at processen med rangordningen er enkel, hvorimod processen med målingen er omstændelig.

Skal man da beregne medianen baseret på et gennemsnit af de to midterste af n observationsværdier eller skal man beregne medianen baseret på gennemsnittet af to medianer, hvor de n observationer er opdelt i to?

$$V\left(Md_n\right)=\frac{\left(2\cdot n-1\right)\cdot\pi_n\cdot\sigma^2}{2\cdot\left(2\cdot n^2+n-2\right)} \qquad V\left(Md_{2\_n/2}\right)=\frac{\left(3\cdot n/2-1\right)\cdot\pi_{n/2}\cdot\sigma^2}{2\cdot2\cdot\left(n/2\right)\cdot\left(3\cdot n/2+1\right)} \qquad Ratio=\frac{V\left(Md_n\right)}{V\left(Md_{2\_n/2}\right)}$$

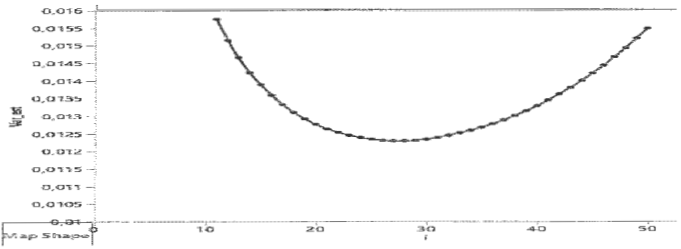


Illustrationen oven for viser, at man bør beregne medianen baseret på et gennemsnit af de to midterste af n observationsværdier fremfor at opdele observationerne i to.

Men når rangordningen er enkel og målingen er omstændelig bør man frasortere en tilfældig af de n observationsværdier. Hvorvidt det er en tilfældig blandt n observationer eller en tilfældig blandt de to midterste observationer gør ingen forskel idet det altid vil være en af de to midterste, der vil blive bragt i anvendelse.

Såfremt man holder fast i at anvende to observationsværdier bør man overveje om det, ved n lige, er anvendelsen af de to midterste observationer, der er den meste efficiente estimator for medianen.

Ved et simulationsstudie med 1 million simulationer, n=100 og $X \sim N(\mu = 0, \sigma^2 = 1)$ tyder det på, at medianen beregnet på grundlag af gennemsnittet af nedre og øvre kvartil er en mere efficient estimator: (27 og 73 procent percentilerne tyder på at være det optimale)



Efficiensen af medianestimator, nedre og øvre kvartil:

$$Md_{Q13} = \frac{Q_1 + Q_3}{2} \text{ og } V\left(Md_{Q13}\right) \approx \frac{\pi \cdot \sigma^2}{2,5 \cdot n} = 1,256 \cdot \frac{\sigma^2}{n}$$

Såfremt medianestimatoren, Md_{Q13}, gennemsnittet af nedre og øvre kvartil, anvendes som estimator for gennemsnittet kan man ud fra ovenstående konkludere, at denne estimator er lige så efficient som $\overline{X}$, blot man har et datagrundlag, der er godt 25 % større og X kan antages at være normalfordelt.

Såfremt man inddrager alle tre kvartiler ved medianestimatet, Md_{Q123}, tyder simulationsstudier på at $Md_{Q123} = \frac{Q_1 + Q_2 + Q_3}{3}$ og $V\left(Md_{Q123}\right) \approx \frac{\pi \cdot \sigma^2}{2,667 \cdot n} = 1,18 \cdot \frac{\sigma^2}{n}$.

Såfremt man inddrog 50 % percentilen samt to andre percentiler ved medianestimatet, Md_{P123}, tyder simulationsstudier på at 18 % og 82 % giver den mest efficiente estimator:

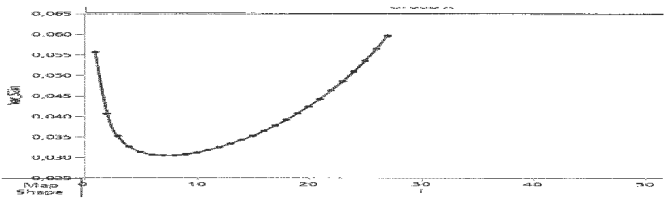
$$Md_{P18} = \frac{P_p + P_{0,50} + P_{1-p}}{3} \text{ og } V\left(Md_{P18}\right) \approx \frac{\pi \cdot \sigma^2}{2,74 \cdot n} = 1,14 \cdot \frac{\sigma^2}{n}.$$

Efficiensvurderinger af alternative variansestimatorer.

Når $X \sim N(\mu,\sigma^2)$, har estimatoren for variansen, S^2, en varians på:

$$V(S^2) = \left(\frac{\sigma^2}{n-1}\right)^2 \cdot V(\chi^2_{n-1}) = \frac{2\cdot\sigma^4}{n-1}, \quad \text{da} \quad \frac{(n-1)\cdot S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{og} \quad V(\chi^2_{n-1}) = 2\cdot(n-1)$$

Ved anvendelse af 2 symmetriske percentiler kan man opnå disse estimater for variansen: $S^2_{i:n} = \left(\frac{X_i - X_{n-i+1}}{2\cdot z_{(i-0,5)/n}}\right)^2$



Den mest efficiente estimator for variansen tyder ved simulationsstudier på at være en nedre percentil i intervallet $\frac{i}{n} \in [0,05;0,10]$ $V(S^2_{0,07}) \approx 3\cdot\frac{\sigma^4}{n}$

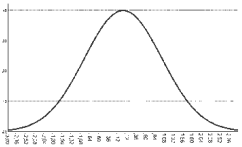
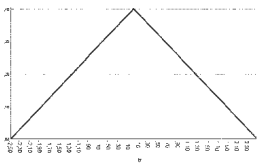
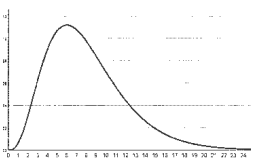
Såfremt variansestimatoren, $S^2_{0,07}$, anvendes som estimator for σ^2 kan man ud fra ovenstående konkludere, at denne estimator er lige så efficient som S^2, blot man har et datagrundlag, der er cirka 50 % større og X kan antages at være normalfordelt.

Afslutning.

Såfremt man vil begrænse sig til to percentiler (X_p og X_{1-p}) til estimation af såvel gennemsnit som varians kunne et kompromis være at anvende 16 %- og 84 %-percentilerne. De tilhørende estimatorer er omkring 10-20 % mindre efficiente end ved de optimale valg af percentiler.

Argumentet for at anvende 16 % percentilen beror på, at det er i dette område, at produktet af de to estimatorers varians er mindst.

Konsekvenser af afvigelser fra normalfordelingsantagelsen illustreret ved n=100 ved 1.000000 simulationer:

Fordeling af X:	$E\left(\hat{\Theta}\right)$ μ $\overline{P_{0,25}}$ $\overline{P_{0,16}}$	$V\left(\hat{\Theta}\right)$ $\sigma^2_{\bar{X}}$ $s^2_{P0,25}$ $s^2_{P0,16}$	$E\left(\hat{\Theta}\right)$ σ^2 $\overline{s^2_{0,07}}$ $\overline{s^2_{0,16}}$	$V\left(\hat{\Theta}\right)$ $\sigma^2_{S^2}$ $s^2_{S^2_{0,07}}$ $s^2_{S^2_{0,16}}$
Z 	**0** 0,00019 0,00013	**0,01000** 0,01235 0,01359	**1,00000** 0,99946 1,00341	**0,02020** 0,03047 0,03653
Triangulær [-2,5;2,5] 	**0** 0,00009 0,00019	**0,01042** 0,01537 0,01535	**1,04167*)** 1,10071 1,18836	***0,01541**)*** 0,02501 0,04864
χ^2_8 	**8,00000** 7,66961 8,03463	**0,16000** 0,18865 0,22317	**16,0000** 15,0352 14,7867	***8,9988**)*** 9,6067 9,6814

$$\overline{P_{0,16}} = \frac{X_{0,16} + X_{0,84}}{2}$$

$$*)\ 2 \cdot \int_0^{2,5} x^2 \cdot \left(\frac{2}{5} - \frac{4}{25} \cdot x\right) dx = 2 \cdot \left[\frac{10 \cdot x^3 - 3 \cdot x^4}{75}\right]_0^{2,5} = 1,04167$$

**) Værdier fra simulationer.

En opsummering af medianestimatoren, hvor 16 %- og 84 %-percentilerne er anvendt:

For normalfordelingen er estimatorerne for middelværdi og for varians forventningsrette. Den relative efficiens af estimatorerne er hhv. 1,359 og 1,808 (0,03653/0,02020).

For den trianguläre fordeling er det kun estimatoren for middelværdi der er forventningsret. Den relative efficiens er 1,473 (0,01535/0,01042). Estimatoren for variansen er upward biased.

For chi-i-anden fordeling med 8 frihedsgrader er estimatoren for middelværdien svagt upward biased. Den relative efficiens er 1,395 (0,22317/0,1600). Estimatoren for variansen er downward biased.

For χ^2-fordelingen med 8 frihedsgrader tyder gennemsnittet af 16%- og 84% percentiler på at være næsten forventningsret.

Ved en nærmere undersøgelse viser det sig, noget overraskende, at:

$$\frac{\chi^2_{v,q} + \chi^2_{v,1-q}}{2} \approx v \quad for \quad q = P(Z > 1) = 0{,}1587 .$$

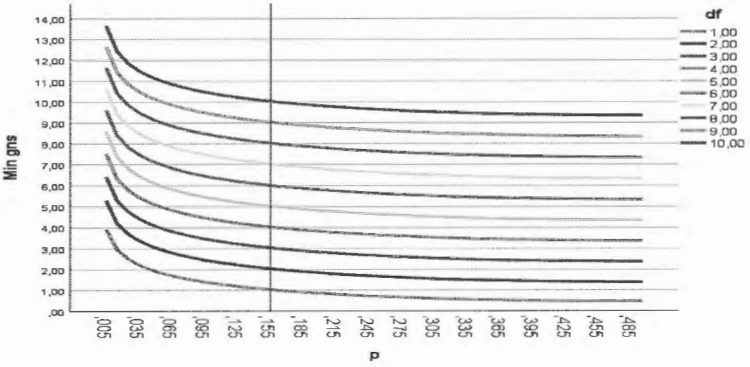
Differencerne for v<5 er i størrelsesorden 0,01.

Nedenfor er resultaterne af simulationsstudier vist og det fremgår, at gennemsnittet af 16%- og 84% percentiler er næsten forventningsret uanset antal frihedsgrader. Med faldene antal frihedsgrader forringes den relative efficiens af estimatoren væsentligt. Estimatoren af variansen er mere udtrykt downward biased ved faldene antal frihedsgrader.

Fordeling af X:	$E(\hat{\Theta})$ μ $\overline{P_{0,16}}$	$V(\hat{\Theta})$ $\sigma^2_{\bar{X}}$ $s^2_{P0,16}$	$E(\hat{\Theta})$ σ^2 $\overline{s^2_{0,16}}$	$V(\hat{\Theta})$ $\sigma^2_{S^2}$ $s^2_{S^2_{0,16}}$
χ^2_8	**8,00** 8,03463	**0,160** 0,22317	**16,00** 14,7867	***8,9988**)*** 9,6814
χ^2_7	**7,00** 7,04	**0,140** 0,197	**14,00** 12,80	***7,3227**)*** 7,42
χ^2_6	**6,00** 6,04	**0,120** 0,170	**12,00** 10,79	***5,7835**)*** 5,54
χ^2_5	**5,00** 5,04	**0,100** 0,143	**10,00** 8,79	***4,4090**)*** 3,85
χ^2_4	**4,00** 4,04	**0,080** 0,115	**8,00** 6,79	***3,2025**)*** 2,45
χ^2_3	**3,00** 3,04	**0,060** 0,087	**6,00** 4,80	***2,1624**)*** 1,36
χ^2_2	**2,00** 2,04	**0,040** 0,060	**4,00** 2,84	***1,2810**)*** 0,60
χ^2_1	**1,00** 1,04	**0,020** 0,032	**2,00** 0,99	***0,5607**)*** 0,13

**) Værdier fra simulationer.

Nedenfor er gennemsnitet af symmetriske percentiler i χ^2-fordelinger med 1 til 10 frihedsgrader illustreret.



Konklusion:

Hvis man anvender to observationsværdier, hvor X er normalfordelt, som estimator for såvel middelværdi som for varians bør man anvende 16 % og 84 % percentilerne.

Estimatorerne er ikke så efficiente som de parametriske estimatorer, men ved en stikprøvestørrelse, der er omkring 80 % større ville man kunne kompensere ift. den parametriske estimator af variansen. Ift. middelværdi ville en 40 % større stikprøve være tilstrækkelig.

Hvis X er skævt fordelt i en form, der matcher en χ^2-fordelingen med et vilkårligt antal frihedsgrader så er estimatoren for middelværdien tæt på at være forventningsret. Estimatoren for variansen er dog downward biased.

Maritz og Jerret (1978) *A Note on Estimating the Variance of the Sample Median.*

Journal of the American Statistical Association, Vol. 73, No. 361 (Mar., 1978), pp.194-196.

Arnold et. al. (2008) *A First Course in Ordered Statistics*. Wiley

Is it possible to reduce smoking by increasing taxes?

Anders Milhøj
Department of Economics
University of Copenhagen
anders.milhoj@econ.ku.dk

In Denmark all ideas to reduce smoking are discussed. It is commonly accepted that increasing prices is an efficient tool to reach this goal. In a Danish context taxes are frequently used for such purposes as we already have a heavy taxation of all tobacco products as we also tax alcohol, fuel and even for some years ago fat in goods like meat, cheese, chocolate etc.

The question is however whether it is true that the smoking could be reduced by increasing taxes. In this paper several statistical time series analyses using procedures - PROC VARMAX, PROC X12, PROC ARIMA and PROC UCM - in the SAS® ETS software package is performed as an attempt to answer this question. The data is series for monthly and yearly sales of cigarettes combined with series for the price of cigarettes and the overall consumer price index together with detailed information on previous changes in the consumer tax for cigarettes.

The results indicate a price elasticity of at least 0.6, but the chock effect of a sudden price increase by an increasing tax fades out rather fast.

Sales of cigarettes in demark

Statistics Denmark publishes data for the sales of cigarettes. The basis for this series is the detailed information upon the tax of cigarettes which is available each month. Moreover the consumer price indices for the price of cigarettes and also the general consumer price index are also published by Statistics Denmark are used. The time series applied in the analyses start January 2001 and data up to December? 2018 is used in this paper - a total of 18??0 observations.

The Danish taxation on cigarettes is heavy. Roughly speaking the tax for a single cigarette at present is 1? Danish Kroner which is say 20 cents. The politicians discuss how much this tax should be increased in order to drastically reduce smoking as an imitative to ensure public health. It is remarkable that the taxation on cigarettes in some European countries is even higher than in Denmark Experiences form e.g. Norway are often referred to as the price of cigarettes in Norway is much higher than in Denmark due to taxes.

The yearly sales and price series for cigarettes is plotted in Figure 1. The general picture is of course the price has increased due to inflation etc. The sales decreased in

the eighties, then increased in the period up to 2005 and afterwards the sales decreased rapidly - almost by a third.

Figure 1 Yearly sales and price series for cigarettes

The price as such is of course not so relevant as inflation, exchange rates etc affect the price in (fixed) Danish Kroner. Instead the relative price index defined as

$$\text{relative price index} = \frac{\text{price of cigarettes}}{\text{consumer price index}}$$

is used. In **Fejl! Henvisningskilde ikke fundet.** the sales series and the series of the relative price of cigarettes are plotted. In Figure 2, moreover, the two series are log-transformed as a starting point for the analyses to come.

Figure 2 Log-transformed sales series and the log-transformed relative price series.

Time series models for the yearly series using PROC VARMAX

It is hard to see from Figure 1 and Figure 2 whether the price series affects the sales series. A thorough analysis following he principles by Box & Jenkins ends up with a rather simple model which tells that the year to year number of sold cigarettes is dependent upon the year to year relative price series with a lag one moving average term included in order to explain for some autocorrelation. This model could be estimated by several SAS ETS® procedures. Here is the code for an analysis using PROC VARMAX.

```
proc varmax data=log_year print=all plots=all;
model log_salg=log_rel_pris/
dif=(log_salg(1) log_rel_pris(1)) method=ml q=1 p=0 noint;
where year(date)>=1977 and year(date)<2019;
id date interval=year;
run;
```

The estimated parameters of the resulting model are

Model Parameter Estimates

Equation	Parameter	Estimate	Standard Error	t Value	Pr > |t|	Variable
log_Salg	**XL0_1_1**	-0.65697	0.20672	-3.18	0.0029	log_rel_pris(t)
	MA1_1_1	0.26786	0.14011	1.91	0.0633	e1(t-1)

which gives the ARIMA(0,1,1) model

$$(1 - B)S_t = -0.66(1 - B)P_t + \varepsilon_t - 0.27\varepsilon_{t-1}$$

The regression parameter, - 0.66, is clearly different from zero, but it is not significantly different from one. This means that a hypothesis of price elasticity for sales of cigarettes equal one is accepted. Or in other words that the budget for purchase of cigarettes on a macro level is fixed to a specific amount. This model passes the usual checks for model fit. However, two outliers for year 2014 and 2018 are present as seen from the Model Plot, Figure 3.

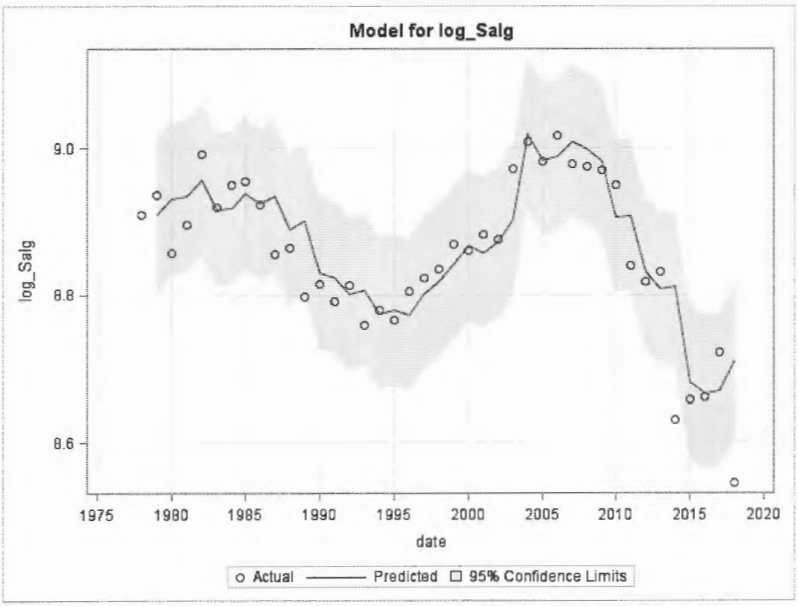


Figure 3 Model plot for the model estimated by PROC VARMAX

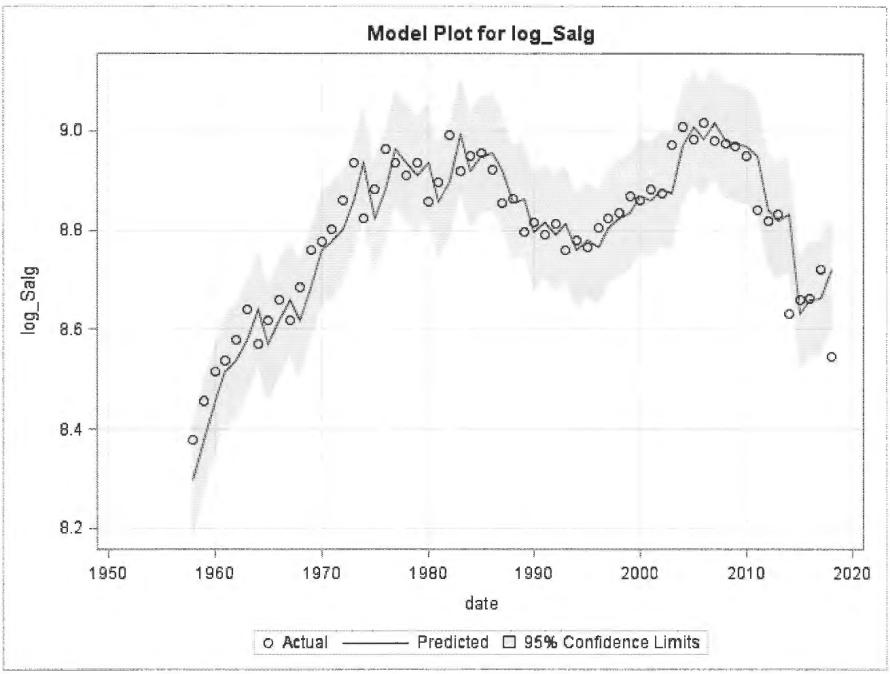
Time series models for the yearly series using PROC UCM

Another approach to analyze these yearly series is to apply unobserved component models by SAS® PROC UCM. In this setup the sales series is decomposed into a level component, a trend component which both could be time varying and an irregular component which could include some autocorrelation parameters. it also possible to include a regression and even to let the regression coefficient vary over time.

The actual series, the log-sales series, has no visible trend, so a trend component is excluded. The level seems however to vary over time so a simple unobserved component models is estimated by this code

```
PROC UCM data=log_year plot=all;
id date interval=year;
model log_salg;
level;
irregular p=1;
estimate extradiffuse=2;
run;
```

The results indicate no serious autocorrelation problems. The resulting Model Plot - Forecast Plot is



This model could be extended by a regression om the relative price series with a randomreg statement as

```
PROC UCM data=log_year plots=all;
id date interval=year;
model log_salg;
level ;
randomreg log_rel_pris;
irregular q=1;
estimate extradiffuse=2;
run;
```

The resulting parameters are

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	**Error Variance**	0.00016568	0.0002981	0.56	0.5783
Irregular	**AR_1**	-0.72486	0.41766	-1.74	0.0826
Level	**Error Variance**	0.00206	0.0007642	2.70	0.0069
log_rel_pris	**Error Variance**	4.51976E-10	4.4797E-7	0.00	0.9992

The results of this table prove that the error variance of the regression component is zero and hence the regression coefficient is constant, the randomreg plot, that is the plot of the time varying regression coefficient is a horizontal line at $\beta = -0.6314$.

A regression with a fixed regression coefficient is fitted using PROC UCM by the code

```
PROC UCM data=log_year plots=all;
id date interval=year;
model log_salg=log_rel_pris;
level ;
irregular p=1;
estimate extradiffuse=2 ;
run;
```

The result again gives a regression coefficient corresponding to a the price elasticity 0.6314.

Final Estimates of the Free Parameters

Component	Parameter	Estimate	Approx Std Error	t Value	Approx Pr > |t|
Irregular	**Error Variance**	0.00016568	0.0002981	0.56	0.5783
Irregular	**AR_1**	-0.72486	0.41762	-1.74	0.0826
Level	**Error Variance**	0.00206	0.0007642	2.70	0.0069
log_rel_pris	**Coefficient**	-0.63139	0.20406	-3.09	0.0020

The monthly series

It is possible to give more precise answers by use of monthly data as many changes in the taxation were effective from say April 1st. and not at the beginning of a new year. Monthly sales data is however only available from January 1. 2001, so it is impossible to go as far back in time as in the analysis by yearly data.

Figure 4 presents the log-transformed relative price index of cigarettes of to the general consumer price index. The red vertical lines represent changes in the taxation on cigarettes.

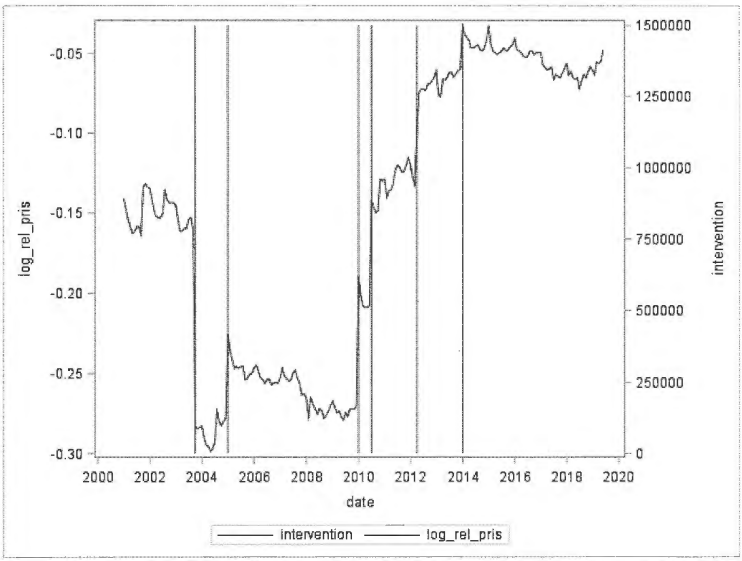


Figure 4 Log-transformed relative price index of cigarettes of to the general consumer price index with marked changes in the taxation on cigarettes.

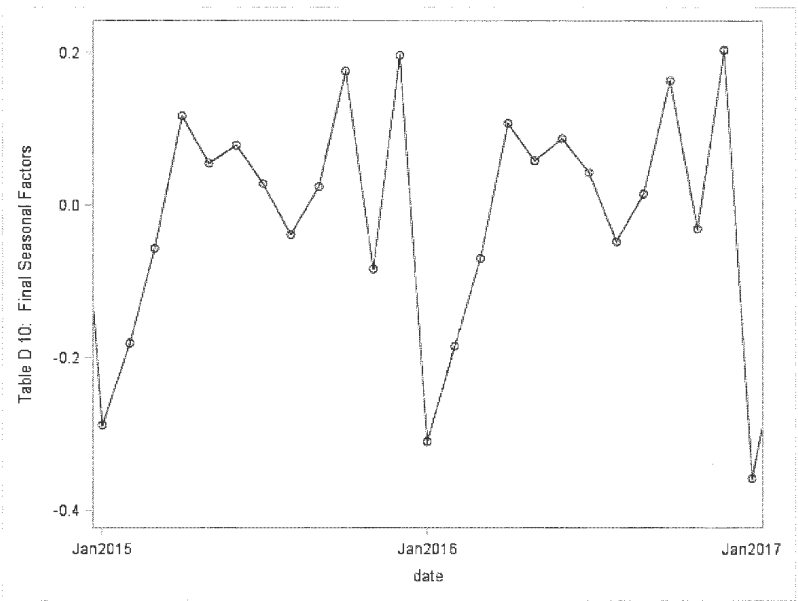
Note that the change in taxation by October 1. 2003 gave a price reduction as the special tax on cigarettes was reduced in order to reduce border trade as limits on private import from abroad were deleted.

Seasonal structure of the monthly series

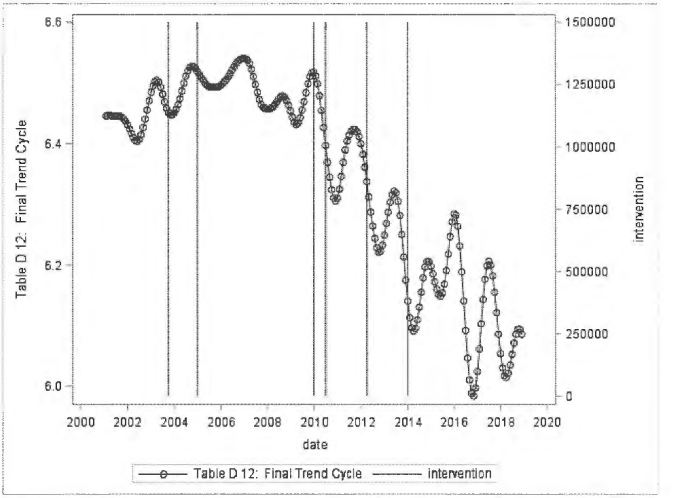
The monthly sales series include some seasonal variation which could be analyzed by the Census Seasonal Adjustment method, which is possible in SAS® by PROC X11, X12 or X13. The code is

```
PROC X12 data=i date=date;
var l_sales_month;
outlier type=ao;
x11 mode=add  trendma=23;
regression predefined=(td Easter(4));
automdl;
output out=out a1 c17 d10 d11 d12 d13 c17;
run;
```

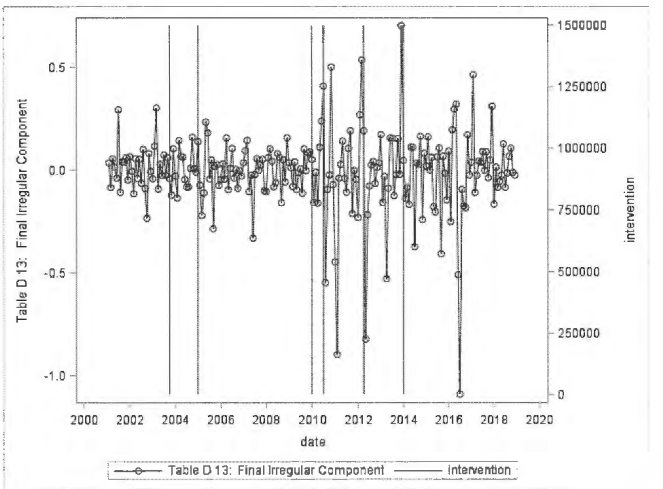
The sales series is decomposed in output dataset into a trend cyclic, a seasonal component and an irregular component. Figure? presents two year of the seasonal component. In this plot the seasonal structure is seen in detail. Most remarkable is the small sales figure in January and high sales in October and December. One reason for this is the holiday season in December.



The next Figure gives the estimated trend cyclic component compared to the changes in taxation. Note that the trend cyclic component in the code is smoothed by a 23 terms Henderson average and not by the default 13 terms average. It is clear that the changes in taxation seem to have some effect on the sales series.



The irregular component compared to the changes in taxation, Figure?, also seems to be affected by changes in the taxation. But large values of the irregular component are also present for many more months.



PROC ARIMA applied to the monthly series

By usual BOX & Jenkins/1976) model identification techniques we end up with an ARIMA(1,0,0)×ARIMA12(0,1,1) model

$$(1 - 0.25)(1 - B^{12})S_t = (1 - 0.72B^{12})\varepsilon_t \text{ , } var(\varepsilon_t) = 0.2584^2$$

This model is extended by the log-transformed relative price seires as an input varialbe by the code

```
PROC ARIMA DATA=tobak.cigaret_month PLOTS=ALL;
IDENTIFY VAR=l_sales_month(12)
crosscor=(log_rel_pris(12));
    ESTIMATE p=(1) Q=(12) input=(log_rel_pris) noint
METHOD=ML;
 run;
```

The estimation results are given by the table

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.88115	0.06384	13.80	<.0001	12	l_sales_month	0
AR1,1	0.15248	0.06710	2.27	0.0231	1	l_sales_month	0
NUM1	-1.44333	0.22174	-6.51	<.0001	0	log_rel_pris	0

The estimate elasticity 1.44 which is only hardly significant different from one confirms the finding in the analyses of the yearly series.

This model could be extended by several intervention components following the principles suggested by Box & Jenkins(1976).

Every change in the taxation was announced before the change was actually put in action. This means that it was possible to buy extra cigarettes before an announced higher taxation and in case of an announced reduced tax, to purchase some extra cigarettes. This point is relevant as the stores of cigarettes in the shops are already taxed, so the sales considered in this analyses is from the wholesaler part to the detail shops. In the analyses this means that an intervention is relevant also the month before the actual new tax is introduced. This intervention is formally defined as

$B_t = 0$ for all t unless $B_t = 1$ for t = the month before the change in taxation.

Also of course the new taxation could lead to an immediate reaction in the sales which also could be due to delayed sales figure in case of a tax reduction or reduced sales in case of a tax increase due to increased sales prior to the actual date of the taxation change. . This intervention is formally defined as

$I_t = 0$ for all t unless $I_t = 1$ for t = the first month sfter the change in taxation.

A consumer reaction the first month after the change in taxation could also indicate changes in smoking behavior due to price changes - at least this is often an argument for the change in taxation. The question is whether this reaction is a lasting effect. . This intervention is formally defined as

$L_t = 0$ for all t, but $L_t = 1$ for t = all months after the change in taxation.

Often the effect of an intervention is however declining - perhaps smokers gets used to the new price levels and find the money for say increasing prices of cigarettes elsewhere in their personal budget. The idea is that the effect is slowly decaying by a factor δ as

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

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. 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).

Models of this kind are easily fitted by PROC ARIMA. The next code gives an application of only one of the six changes in taxation in the period of analysis. In this first application only the change in taxation by January 1. 2012 is included. This intervention is modeled by an effect the month before and an exponentially declining effect afterwards. It however turns out that the effect i January 2012 was very small, while the effect form February 2012 and onwards was significant, so the shift by one month as coded by **1** $ tell that the effect began February.

In this model the effect of the increasing tax in January 2012 fades out by the damping factor $\delta = 0.25$ which in fact is insignificant meaning that the smokers reacted with some kind of chock but rapidly got used to the new price level.

```
PROC ARIMA DATA=tobak.cigaret_month PLOTS=ALL;
IDENTIFY VAR=l_sales_month(12)
  crosscor=(int_2012(12) lead_int_2012(12));
```

```
ESTIMATE p=(1) Q=(12) METHOD=ML noint
  input=( 1 $/(1) int_2012  lead_int_2012);
*OUTLIER TYPE=(AO LS) ID=DATE;
RUN;
quit;
```

The results are given by the table

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.70166	0.05810	12.08	<.0001	12	l_sales_month	0
AR1,1	0.23695	0.06938	3.42	0.0006	1	l_sales_month	0
NUM1	-0.90612	0.22612	-4.01	<.0001	0	int_2012	1
DEN1,1	0.25191	0.23980	1.05	0.2935	1	int_2012	1
NUM2	0.46619	0.22110	2.11	0.0350	0	lead_int_2012	0

Variance Estimate 0.061838

Std Error Estimate 0.248673

When this model is extended by the log-transformed relative price series the code becomes

```
PROC ARIMA DATA=tobak.cigaret_month PLOTS=ALL;
IDENTIFY VAR=l_sales_month(12)
  crosscor=(log_rel_pris(12) int_2012(12)
lead_int_2012(12));
ESTIMATE p=(1) Q=(12) METHOD=ML noint
  input=(log_rel_pris 1 $/(1) int_2012  lead_int_2012);
*OUTLIER TYPE=(AO LS) ID=DATE;
RUN;
quit;
```

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.84909	0.05866	14.47	<.0001	12	l_sales_month	0
AR1,1	0.14641	0.06907	2.12	0.0340	1	l_sales_month	0
NUM1	-1.34565	0.22566	-5.96	<.0001	0	log_rel_pris	0
NUM2	-0.77058	0.22030	-3.50	0.0005	0	int_2012	1
DEN1,1	0.12638	0.28309	0.45	0.6553	1	int_2012	1
NUM3	0.53645	0.21805	2.46	0.0139	0	lead_int_2012	0

Variance Estimate	0.054976
Std Error Estimate	0.23447
AIC	8.428287
SBC	28.30752
Number of Residuals	203

If the reaction to the tax increase is supposed to be everlasting the code is

```
PROC ARIMA DATA=i PLOTS=ALL;
IDENTIFY VAR=l_sales_month(12)
     crosscor=(log_rel_pris(12) lead_int_2012(12)
level_2012(12));
ESTIMATE p=(1) Q=(12) METHOD=ML noint
     input=( log_rel_pris 1 $ level_2012  lead_int_2012);
OUTLIER TYPE=(AO LS) ID=DATE;
RUN;
```

and the estimation results become

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.92456	0.08256	11.20	<.0001	12	l_sales_month	0
AR1,1	0.10362	0.06759	1.53	0.1253	1	l_sales_month	0
NUM1	-0.63115	0.33001	-1.91	0.0558	0	log_rel_pris	0
NUM2	-0.20200	0.06240	-3.24	0.0012	0	level_2012	1
NUM3	0.49397	0.21847	2.26	0.0238	0	lead_int_2012	0

Variance Estimate 0.053383

Std Error Estimate 0.231048

The price elasticity is in this model estimated as 0.63 and not significantly different from one. The effect of the tax change the month before, that is by December 2011, was an increasing sales figure where the estimate 0.49 corresponds to an increase by $(e^{0.49} - 1)\times 100\% = ??\%$. The estimate -0.20 for the level shift input variable tells that the increasing tax reduced the sales by $(1- e^{-020})\times 100\% = ??\%$ all months from February 2012 and afterward.

References

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

Recent Developments in Danish Inequality

Mette Suder Franck

This paper is based on Danish register data describing individual-level income between 1984 and 2016 and argues that whether focusing on the cross-sectional or panel data dimension matters crucially for the resulting growth statistics.

When investigating the cross-sectional dimension - as most international studies are usually limited by their data to do - the analysis shows that real incomes have increased by far more for the top parts of the distributions. When instead utilizing the panel dimension the pattern is reversed, such that it is the individuals who had the lowest incomes in 1984, who are the ones that experienced the highest income growth rates during the 32-year period.

Together the two analyses compliment each other and offer important information for policymakers, as they suggest that although inequality has increased between 1984 and 2016 a person's relative placement in the income distribution may not persist throughout his lifetime.

Introduction

Though economic inequality is usually analyzed as a separate subject it is hypothesized to have important consequences for several other societal outcomes. For instance by Stiglitz (2013), who argues when describing the recent rise in American inequality that this development has resulted in "*[...] an economic system that is less stable and less efficient, with less growth, and a democracy that has been put into peril.*" (p. xli).

Stiglitz bases his claim on an analysis of the situation in the United States but his concerns are not uniquely American. A similar argument have been formulated by Thomas Piketty (2014) and serves as a central motivation behind his bestselling book "Capital in the Twenty-First Century" that sparked an interest in economic inequality that reached far beyond that of his fellow economists'.

The present paper contributes to the growing body of inequality literature spurred by Piketty's work with an analysis of recent developments in Danish economic inequality. With a focus on the adult Danish population during the period from 1984 to 2016, the

study aims at describing the development in inequality by analyzing pretax and posttax incomes for different groups of the income distributions before and after taxes.

The main findings confirm the tendency of increased inequality reported in other recent studies and show that the top parts of the income distributions have by far experienced the highest growth rates during the period, hereby corroborating the proposition of dissimilar growth experiences across the income distribution and the relevance of computing disaggregated growth statistics emphasized by Piketty et al. (2018).

The paper also adds to the knowledge about Danish inequality with an analysis of the distributions of total income between men and women. This investigation reveals that despite a tendency towards more equality, the distribution of income between the two genders was still highly unequal in 2016.

The paper further contributes to the literature on inequality by comparing the main results concerning the cross-sectional growth experiences with analyses based on a balanced panel dataset. The two approaches differ in that the former presents the development in the income distributions, whereas the latter describes the income growths of the same individuals over time. This comparison shows that when investigating the incomes of the same individuals over time the picture of inequality is strikingly different, as the income growth of the initial bottom income group is now notably higher than that of other parts of the distribution.

This finding suggests that having access to panel data allows for a fuller picture of inequality and a more solid empirical foundation for policymakers who wish to address the issue of income distribution.

Empirics and Inequality

Not before Simon Kuznets' work in the beginning and middle of the twentieth century did statistics and empirical analyses become a central part of the economic literature on inequality.

Kuznets based his highly influential theory on analyses of data on the economic development in the United States between 1913 and 1948 and argued that inequality will develop according to a bell-shaped curve as the economy grows (Kuznets, 1955).

The importance of the theory's empirical foundation have later been stressed by Thomas Piketty, who argues that the influence of the 'Kuznets curve' on the inequality literature can not be separated from the fact that its thorough empirical basis offered both a statistical source and a method to use for evaluation of the theory of inequality (Piketty,

2014). However, as the statistical foundation was limited in both its dimension of time and space the issue of the theory's generality was by the 1980's beginning to get seriously questioned.

Milanovic (2016) argues that one reason why Kuznets' hypothesis was able to dominate the field for so many years despite its empirical challenges, was the lack of a coherent alternative. At least this was the case until Piketty published "Capital in the Twenty-First Century" in 2013. One of Piketty's main conclusions, which he bases on an extensive collection of historical data, is that the dynamics of the income distribution involves both forces of convergence and divergence and that there are no underlying mechanisms that will naturally ensure that we end up with the former rather than the latter.

The book became a somewhat surprising bestseller and spurred an interest in empirically studying inequality by considering the full distribution of income instead of merely relying on scalar numerical representations, such as the Gini coefficient.

Despite its popularity as a tool for assessing inequality, the Gini coefficient has the disadvantage compared to Piketty's approach of being particularly sensitive to the middle of distributions. As a consequence, the Gini coefficient may fail to properly capture the distributions' tails and subsequently underestimate of the income inequality.

Danish Inequality

Irrespective of methodological approaches newer studies of Danish income inequality generally agree that inequality has increased in recent years but that it remains low compared to other developed economies (as concluded by e.g. De Økonomiske Råd (2016) and Causa et al. (2016)).

Quitzau (2017) characterizes the income distribution according to a range of descriptive statistics and briefly reflects on the income differences between men and women - slightly more women at the bottom and significantly fewer at the top. His insights are based on the distribution of total income, which distinguishes Quitzau's analysis from the majority of studies concerned with gender income inequality, as these usually apply a narrow focus on describing discrepancies in labor income.

Quitzau (2017) focuses on providing an in-depth analysis of the features characterizing the Danish income distribution in 2014 and not on describing the development in the income distribution over time.

In Atkinson and Bourguignon (2015) on the other hand, the time dimension receives considerable attention. Inspired by Piketty, the authors investigate the long run develop-

ment in Danish inequality and combine different data sources to describe the period from 1870-2010. This investigation leads them to conclude that the long run development has entailed a notable decrease in the top income shares, while there are indications of a rise in these shares towards the end of the period.

The inequality investigation conducted in Atkinson and Bourguignon (2015) is in many ways similar to the one in this study, but does not investigate the growth rates for different parts of the income distribution. Moreover, as is also the case for other recent studies of Danish inequality, such as De Økonomiske Råd (2016), they do not utilize the panel dimension of the data to investigate how the incomes of the same set of individuals change over the years.

In the present study, both analyses exploiting the data's panel dimension and investigating the potential heterogeneity in growth experiences across the income distributions are conducted.

Empirical Approach

To describe the development in Danish income inequality, I investigate the income distribution during the years from 1984 to 2016 based on Danish register data and limit the benchmark sample to include individuals above 19 years of age.

To check the sensitivity of the results to changes in the baseline sample, I have computed growth statistics for samples including 15 to 19-year-olds and excluding negative incomes. In both cases, the alternative specifications indicate a bigger increase in inequality during the 32-year period compared to the baseline sample. This suggests that the paper's findings may be conservative estimates of actual development in Danish inequality between 1984 and 2016 [1].

The data contains a politically induced data break, due to the decision of grossing up all social transfers beginning in 1994, which will affect the pattern observed for the pretax income distributions over the years. In my investigation, I do not attempt to correct for the grossing up on pretax income but merely try to provide an indication of the effect by including the income statistics for 1994 to facilitate separate analyses of the periods before and after 1994 [2].

The study's empirical approach relies on presenting descriptive statistics with little

[1] The tables based on alternative specifications are available upon request.

[2] Atkinson and Bourguignon (2015) have conducted an analysis that corrects for these effects and reach conclusions that are qualitatively similar to the ones presented in this paper.

need for any assumptions to be made with respect to the method of computations. Therefore, uncertainties concerning the validity of the results are mainly connected to the data and the choices made in relation to selection of the sample.

Data related issues can affect the study's distributional analyses insofar that there is a tendency of the data to inaccurately capture the income of only certain income groups.

A recent study by Alstadsæter et al. (2017) presents convincing evidence of a higher prevalence of tax evasion among the rich. Since the register data does not cover tax evasion, this finding could mean that they capture a disproportionate share of income across the income distributions. As a result, one would tend to underestimate the incomes at the top of the distribution and consequently also the level of income inequality. It is difficult to evaluate the quantitative effect that the issue of tax evasion might have on the analyses made in this paper. However, given that the implications of Alstadsæter et al. (2017) suggests a downward direction of the potential bias, the reported findings concerning inequality could potentially be stronger than indicated by this study's analyses.

Methodology

The present paper is inspired by the analysis of the American distributional national accounts (DINA) conducted in Piketty et al. (2018), who developed the DINA method as a way to perform in-depth analyses of inequality based on an integration of micro- and macro-based data.

This study does not attempt to construct the Danish DINA, but bases the method for analyzing the development in Danish inequality over time on the approach used in Piketty et al. (2018). Although the computation of the Danish DINA could be interesting for the sake of cross-country comparison, the integration of micro- and macro-level data is arguably less relevant in Denmark than in other countries, due to the unique quality of the register data in terms of depth and coverage (as noted by Esping-Andersen (2017))[3].

The paper analyzes pretax and posttax income inequality for different subgroups delimited by the individuals' positions in the income distributions before and after taxes.

In order to describe the changes in the income distribution I rely on the micro-level information and focus on different subgroups defined according to fractiles: The bottom 50%, the middle 50% - 90%, top 90-99%, top 99%-99.9% and top 99.9%-100%. For each

[3]A comparison of the micro-total, as measured by the sum of pretax income, with the macro-total of national income shows that on average the micro data captures 93% of the macro data. Compared to the American case where the number is only about 60% this is quite reassuring in terms of relying on micro data to describe Danish inequality. The graphs showing these results are available upon request.

of these subgroups, I compute the relative growth rates and average yearly growth rates in order to describe their growth experiences during the full 32-year period from 1984-2016, as well as in the two subperiods from 1984-1993 and 1994-2016.

To construct the growth statistics, I compute each individual's fractional rank in both the pretax and posttax income distribution and then place them in a pretax and postax subgroup based on their real income in both the first and last year of the period[4]. I then calculate the average income within each of the subgroups, which allows for computation of the relative growth rates and average yearly growth rates of the average incomes within each group.

To investigate if men and women have experienced different growths in income, I conduct a similar analysis based on sub-datasets of men and women.

Findings

Table 1 shows the distribution of real pretax and posttax personal income in 1984, 1994 and 2016. As a first point, I want to draw attention to the heterogeneity of growth statistics across the different income groups. This feature would not be captured by analyses focusing solely on population averages, such as the growth in overall macro indicators as GDP per capita or national income per capita.

Keeping this observation of heterogeneity in mind, the next subsections analyze the growth experiences of the different income groups.

Inequality and Development in the Income Shares

To evaluate the development in inequality during the 32-year period, perhaps the most relevant statistics in table 1 is that describing the income shares.

The table shows that the income shares of the top part of the income distribution has increased between 1984 and 2016. This increase in the top shares has been present for both the distribution of the pretax and posttax income, while only the posttax share of the bottom half of the distribution has decreased during the same period. The lack of any noticeable changes in the bottom pretax income shares could be due to the grossing up of the public transfers, as this policy change is arguably likely to have a larger positive effect on the incomes at the bottom of the distribution, given that public transfers makes up a larger share of their total income (as showed by Quitzau (2017)).

[4]Incomes are deflated using the 2015 consumer price index constructed by Statistics Denmark

Table 1: Descriptive Statistics, 1984-2016

1984		Posttax income			Prettax income		
	Number of observations	Average income	Income Threshold	Income share	Average income	Income Threshold	Income share
Full population	3.733.311	135.995		100%	246.197		100%
Bottom 50%	1.866.669	80.743		29,97%	119.555		24,28%
Middle 50%-90%	1.493.311	170.348	138.233	50,10%	307.587	224.389	49,97%
Top 90%-99%	335.998	251.944	215.510	16,67%	548.031	431.751	20,03%
Top 99%-99,9%	33.599	416.779	344.726	2,76%	1.155.765	870.809	4,22%
Top 99.9-100%	3.734	1.058.665	629.665	0,78%	3.659.872	2.028.375	1,48%

1994		Posttax income			Prettax income		
	Number of observations	Average income	Income Threshold	Income share	Average income	Income Threshold	Income share
Full population	4.035.412	158.702		100%	282.209		100%
Bottom 50%	2.017.742	99.990		31,50%	153.885		27,26%
Middle 50%-90%	1.614.136	189.117	153.734	47,66%	335.235	248.293	47,51%
Top 90%-99%	363.180	289.590	240.931	16,40%	604.898	468.649	19,29%
Top 99%-99,9%	36.318	566.276	430.074	3,21%	1.340.812	992.972	4,27%
Top 99.9-100%	4.036	1.901.456	993.868	1,20%	4.665.065	2.396.216	1,65%

2016		Posttax income			Prettax income		
	Number of observations	Average income	Income Threshold	Income share	Average income	Income Threshold	Income share
Full population	4.375.900	214.471		100%	330.809		100%
Bottom 50%	2.187.951	114.181		26,61%	160.474		24,25%
Middle 50%-90%	1.750.359	256.328	190.990	47,80%	394.852	280.723	47,74%
Top 90%-99%	393.831	448.400	355.438	18,81%	749.818	569.381	20,39%
Top 99%-99,9%	39.383	1.077.657	737.839	4,50%	1.891.247	1.297.092	5,10%
Top 99.9-100%	4.376	4.794.670	2.281.437	2,20%	8.126.904	3.903.335	2,46%

Source: Own calculations using register data. Incomes expressed in 2015-prices using CPI from Statistikbauken.dk (PRIS112).

A comparison of the pretax and posttax income shares suggests that the taxes has had a redistributing effect, such that the half of the population with the lowest incomes has had a larger share of total income after taxes than before, while those at the top of the distribution has had a posttax share that is smaller than their pretax share.

Looking at the middle of the income distribution, the 50% to 90% group, the differences between the posttax and pretax income shares are in all three years negligible, which suggests that the redistribution through taxes has mainly affected the income shares of the top 10% and bottom 50%.

Together these developments suggest that inequality has risen slightly between 1984 and 2016 and corresponds to other recent empirical studies of the development in Danish inequality over time, such as De Økonomiske Råd (2016).

Income Growth

Piketty et al. (2018) report that the bottom 50% of the American population has not experienced any pretax income growth between 1980 and 2014, even though the American economy, as expressed by the national income, grew by 61% during the same period.

Inspection of the growth statistics in table 2 shows that this has not been the case in Denmark. All of the five different subgroups have had an increase in real income both before and after taxes between 1984 and 2016, as well as during the two subperiods from 1984-1993 and 1994-2016.

Table 2: Growth Rates, 1984-2016

Relative Growth

	Posttax Income Growth			Pretax Income Growth		
	1984-1993	1994-2016	1984-2016	1984-1993	1994-2016	1984-2016
Full population	13,3%	35,1%	57,7%	12,7%	17,2%	34,4%
Bottom 50%	19,4%	14,2%	41,4%	20,7%	4,3%	34,2%
Middle 50%-90%	8,0%	35,5%	50,5%	8,5%	17,8%	28,4%
Top 90%-99%	11,4%	54,8%	78,0%	10,2%	24,0%	36,8%
Top 99%-99,9%	29,6%	90,3%	158,6%	17,0%	41,1%	63,6%
Top 99.9-100%	106,7%	152,2%	352,9%	46,2%	74,2%	122,1%

Average Annual Growth (in %)

	Posttax Income Growth			Pretax Income Growth		
	1984-1993	1994-2016	1984-2016	1984-1993	1994-2016	1984-2016
Full population	1,4	1,4	1,4	1,3	0,7	0,9
Bottom 50%	2,0	0,6	1,1	2,1	0,2	0,9
Middle 50%-90%	0,9	1,4	1,3	0,9	0,7	0,8
Top 90%-99%	1,2	2,0	1,8	1,1	1,0	1,0
Top 99%-99,9%	2,9	3,0	3,0	1,8	1,6	1,6
Top 99.9-100%	8,4	4,3	4,8	4,3	2,6	2,5

Source: Own calculations using register data. Real growth rates based on CPI from Statistikbanken.dk (PRIS112).

In the full 32-year period the pretax income of the population's bottom 50% grew by 34,4%. The pretax growth rates during the two subperiods from 1984-1993 and 1994-2016 were 20,7% and 4,3%, respectively. This comparison reveals that the growth in the income of the bottom half of the income distribution has mainly taken place during the period's first 10 years.

In comparison to the population average, the development in the incomes of the bottom half of the population was virtually the same for the pretax income, but 16,3 percentage points lower for the posttax income. Interestingly, also the average income among those in the middle of the income distribution was below population average income growth in

both pretax and posttax income, which implies that a large part of the average growth rates for the full population can be attributed to the growth in the incomes of the top 10%. Indeed, the growth rates of the three subgroups making up the top 10% of the population do far surpass those of the population average, just like the average annual growth rates for the top of the income distributions are substantially higher than for the bottom 90%.

Table 2 also shows that the growth rates increase the higher up one moves in the income distributions and that the same tendency of increasing inequality is evident when looking at the average annual growth rates.

As for the development in the income shares, the different patterns of growth among the five income subgroups supports the conclusion of an increase in inequality during the 32-year period.

Inequality and Gender

To investigate the development in gender income differences over time, figure 1 depicts the composition of men and women in each of the income subgroups in 1984, 1994 and 2016 for the benchmark population of adults above the age of 19.

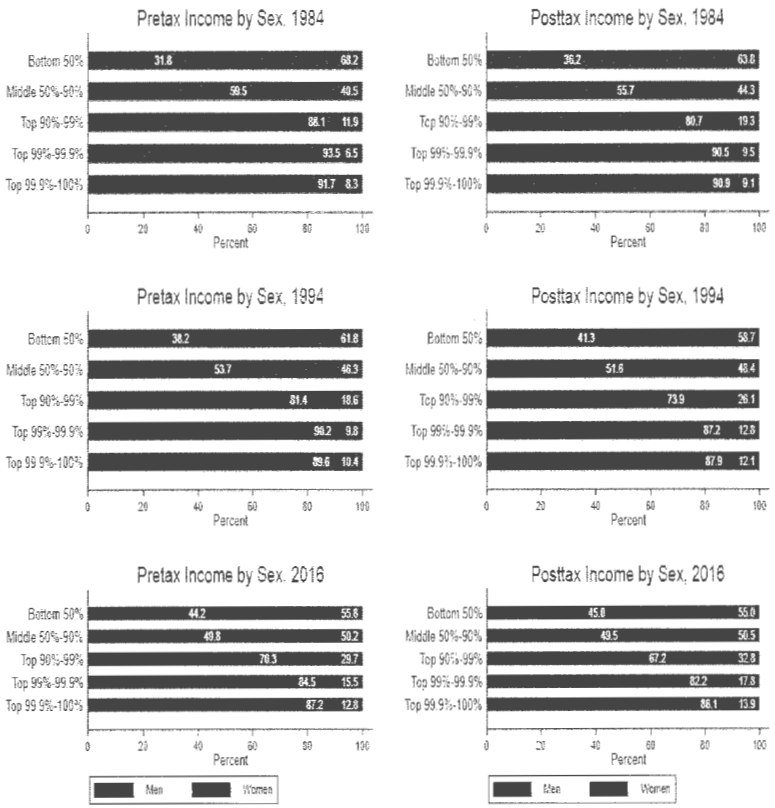
For both pretax and posttax income it is clear that there are noticeable differences in the representation of men and women across the income distribution. In all years, the pretax and posttax income shares of women get lower the higher up one moves in the distribution, albeit those with the very top incomes, the 99.9th to 100th quantile, in few cases deviate from this pattern.

The figure further suggests that redistribution works to mitigate gender income inequality, as differences in the posttax incomes are smaller than those in pretax income, even though posttax incomes in all years are only slightly more equitable.

In an international comparison, the Danish development follows the same pattern as the gender differences reported using the DINA method in France and the Unites States, with high, but slowly decreasing inequality. In addition, the share of women at the very top of the distribution is also quite similar in size, even though the shares reported in the DINA studies are based solely on pretax *labor* income.

In sum, figure 1 shows that men in all years are over-represented at the top of the income distributions despite a tendency towards greater gender equality.

Figure 1: Income Distribution by Gender, 1984-2016



Source: Own calculations based on register data

Income Mobility

Up until now the empirical investigation has been based on the cross-sectional dimension of the data. As a consequence, the previous findings reflect the growth experiences of different parts of the income distribution over time and not the experiences of the same individuals.

In this section, I exploit the panel dimension and analyze the extent of *intragenerational*

mobility, i.e. the mobility of the same individuals during their lifetime (Jäntti and Jenkins, 2015). Such analysis is interesting in relation to an inequality study, because it provides an indication of the degree to which current income matters for future income.

Table 3 summarizes the growth experiences for individuals that belonged to a specific subgroup of the pretax and posttax income distribution in 1984. The statistics are computed based on balanced samples of individuals that were present in both the first and last time period. As individuals might disappear from the sample over the years the number of observations used to calculate the different growth rates and average annual growth rates varies for the periods considered.

Table 3: Growth Rates by 1984 Income Group, 1984-2016

Relative growth

	Posttax Income Growth			Pretax Income Growth		
	1984-1993	1994-2016	1984-2016	1984-1993	1994-2016	1984-2016
Full population	12,1%	44,2%	56,1%	11,1%	22,9%	25,4%
Bottom 50%	48,5%	101,9%	144,5%	46,4%	87,4%	120,5%
Middle 50%-90%	0,9%	27,8%	34,8%	2,5%	11,1%	12,0%
Top 90%-99%	-4,8%	23,0%	20,3%	-0,2%	-0,8%	-8,3%
Top 99%-99,9%	-11,2%	2,8%	12,6%	-3,1%	-17,1%	-17,6%
Top 99.9-100%	-32,0%	-15,3%	9,1%	-10,9%	-23,2%	-15,4%

Average relative growth

	Posttax Income Growth			Pretax Income Growth		
	1984-1993	1994-2016	1984-2016	1984-1993	1994-2016	1984-2016
Full population	1,3	1,7	1,4	1,2	0,9	0,7
Bottom 50%	4,5	3,2	2,8	4,3	2,9	2,5
Middle 50%-90%	0,1	1,1	0,9	0,3	0,5	0,4
Top 90%-99%	-0,5	0,9	0,6	0,0	0,0	-0,3
Top 99%-99,9%	-1,3	0,1	0,4	-0,3	-0,8	-0,6
Top 99.9-100%	-4,2	-0,8	0,3	-1,3	-1,2	-0,5

Source: Own calculations using register data. Incomes expressed in 2015-prices using CPI from Statistikbanken.dk (PRIS112).

The table shows that those who had the lowest real incomes in 1984 are also the ones who experienced the highest growth rates in their incomes. In fact, the table indicates that the growth rates and the average relative growth rates are falling in initial income, such that those who in 1984 had the highest incomes also experienced the lowest income growth. In several cases, the average real incomes have even declined, especially when considering those who had the highest pretax incomes in 1984.

It is reassuring that these conclusions do also apply to the shorter sub-periods, which suggests that the observed patterns do not only reflect changes in income related to the

life cycle such as retirement or being a student.

It is interesting to compare this table to table 2 that contained the growth statistics for the five subgroups of the pretax and posttax income distributions and thus summarized the growth experiences of different parts of the distributions, rather than for specific individuals.

This comparison shows that whether one is calculating the growth statistics for the same individuals or for specific parts of the distribution clearly matters: The resulting growth experiences are actually reversed!

Together the two tables indicate that during the period from 1984 to 2016 the difference between those at the top and those at the bottom of the real pretax and posttax income distributions has gotten larger. However, at the same time those who were at the bottom half of the distributions in 1984 experienced substantially higher income growth than those at the top. The fact that the growth experiences reported in table 3 are heterogeneous suggests that a person's relative position in the income distribution may very well change over time.

Conclusion

Nowadays, the subject of inequality receives wide attention and constitutes the focal point of many empirical studies, which generally agree that within-country inequality has increased in developed economies during recent years. This rise has, in addition to more traditional claims of a relation to economic growth, also been hypothesized to exert an influence on other societal outcomes, such as politics and social cohesion.

This paper adds to the literature on inequality, with an analysis of Danish income inequality between 1984 and 2016 and reports findings that support the claim made by Piketty et al. (2018) of a need for disaggregated growth statistics in inequality analyses. The study utilizes the Danish register data to analyze the development in inequality and confirms the tendency of rising inequality found in other studies. Moreover, it adds novel insights concerning gender inequality and intragenerational mobility to the knowledge about Danish inequality.

The analysis of gender inequality described the distribution of men and women in different subgroups of the income distribution and indicated that in spite of a tendency towards a more equal distribution of total income between the two genders, the share of women has in all years been markedly lower in the top part of the income distributions.

The panel dimension of the register data is unique in that it allows to investigate

how individuals' incomes change during their lifetime. Such analysis of intragenerational income mobility is interesting for inequality studies, as it can help to evaluate the persistence of inequality. If inequality is transient, as indicated by high intragenerational income mobility, it might change the scope and motivation for political interventions aimed at reducing inequality, compared to a situation where mobility is low and inequality therefore highly persistent.

In stark contrast to the paper's main results, the analysis focusing on the growth experiences of the same individuals over time showed that the persons with the lowest incomes in 1984 were also the ones that experienced the largest growth rates during the subsequent 32-year period, hereby suggesting that individuals are likely to change their relative positions in the income distribution over time.

Throughout the study, the paper's methodological approach has been to focus on providing descriptions of Danish inequality between 1984 and 2016 without trying to provide causal explanations of the observed developments. Therefore, a focus on the factors underlying these developments would be an interesting starting point for future studies of Danish inequality.

References

Alstadsæter, A., Johannesen, N., and Zucman, G. (2017). 'Tax evasion and inequality'. Working Paper 23772, National Bureau of Economic Research.

Atkinson, A. B. and Bourguignon, F. (2015). 'Introduction: Income distribution today'. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2, pages xvii – lxiv. Elsevier.

Causa, O., Hermansen, M., Ruiz, N., Klein, C., and Smidova, Z. (2016). 'Inequality in Denmark through the looking glass'. Working Paper 1341.

Esping-Andersen, G. (2017). 'Kapitel 8 - Multidimensional omfordeling i velfærdsstater'. In Ploug, N., editor, *Økonomisk ulighed i Danmark*, pages 147–164. Jurist- og Økonomiforbundets Forlag.

Jäntti, M. and Jenkins, S. P. (2015). 'Chapter 10 - Income Mobility'. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2, pages 807 – 935. Elsevier.

Kuznets, S. (1955). 'Economic growth and income inequality'. *American Economic Review*, 45(1):1.

Milanovic, B. (2016). *Global Inequality: A New Approach for the Age of Globalization*. Harvard University Press.

Piketty, T. (2014). *Kapitalen i det 21. århundrede*. Gyldendal, København, 1. edition.

Piketty, T., Saez, E., and Zucman, G. (2018). 'Distributional National Accounts: Methods and Estimates for the United States*'. *The Quarterly Journal of Economics*, 133(2):553–609.

Quitzau, J. (2017). 'Kapitel 2 - Indkomstfordelingen i Daumark'. In Ploug, N., editor, *Økonomisk ulighed i Danmark*, pages 33–61. Jurist- og Økonomiforbundets Forlag.

Stiglitz, J. E. (2013). *The price of inequality*. W.W. Norton & Co, 1 edition.

Økonomiske Råd, D. (2016). *Dansk økonomi efteråret 2016*. De Økonomiske Råd.

Statistik på økonomiuddannelserne: Forandring fryder - eller?

Nils Karl Sørensen, nks@sam.sdu.dk Institut for Virksomhedsledelse og Økonomi, SDU

Resumé

Gennem de seneste årtier har basiskurserne i statistik på økonomiuddannelserne undergået væsentlige forandringer. Fremkomsten af forbedrede redskaber til at beregne forskellige statistikker er blevet forbedret. Dette har flyttet fokus i formidlingen. Den mere forenklede tilgang til beregningerne har betydet, at vægtningen af kurserne i statistik er blevet drastisk reduceret.

Med udgangspunkt i forfatterens næsten 40 års viden om statistikkurser undersøges denne problemstilling. Der lægges ud en perspektivering af statistikken som metodefag. Dernæst følger en diskussion af kursets typiske temaer i relation til, hvad studenterne skønnes at måtte have af behov. Endelig diskuteres relevansen af undervisningen i statistik på økonomiuddannelserne, og lyset rettes mod, hvad en fremtidig vision kan være for statistikkens rolle.

Diskussionen ledsages af eksempler fra dengang og nu, med perspektiv på læringsprocessen.

1. Baggrund og metodiske tilgange

Empirisk statistik er en videnskabelig metode, der baserer sig på anvendelse af numeriske tal. Statistikkens rolle på økonomiuddannelserne har været som et metodisk redskab for at kunne bidrage til besvarelse af givne problemstillinger. Den anvendte statistiske tilgang øver således indflydelse på de konklusioner, der drages.

Ofte vil man bearbejde materiale fra statistikbanker eller spørgeskemaundersøgelser for at bekræftet eller forkastet hypoteser, der tager sit udspring for eksempel i økonomisk teori.

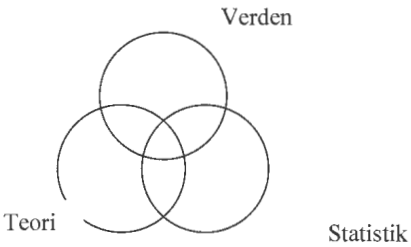
Analytiske arbejdsprocesser tager sit udspring i følgende forhold:

- Virkeligheden (som den observeres)
- Teorien (baseret på observationer formuleres teorier)
- Statistikken (baseret på registreringer af observationer)

Det vil sige, at man indledningsvis konstaterer en forundring. På denne baggrund formuleres en teoretisk økonomisk referenceramme. Det kan eksempelvis være en konsumentadfærd. Denne forundring leder til det teoretisk formulerede forhold, at

konsumenten maksimerer sin nytte ved en given kombination af varer under en budgetbegrænsning. Endelig leder disse overvejelser til opstillingen og udførslen af en spørgeskemaundersøgelse. Her anvendes en repræsentativ stikprøve, og der stilles nogle spørgsmål i form af en række hypoteser, der be- eller afkræftes. For at dette skal kunne lede til konsistente svar, skal den underliggende datagenererende proces gerne være normalfordelt. Hvis dette ikke er tilfældet, skal der anvendes en ikke-parametrisk tilgang for at kunne udlede konsistente udfald af hypoteserne. Allerede på dette tidspunkt kan det konstateres, at den statistiske tilgang let bliver ganske kompleks og antallet af mulige fejlkilder bliver stigende.

Sammenhængen mellem de virkeligheden, den økonomiske teori og den statistisk metodiske tilgang er søgt anskueliggjort i diagrammet:



Gode analyser indeholder elementer fra alle forholdene, og vil derfor finde sted i den lille fællesmængde mellem de tre mængder. Der er således mange fejlkilder! Observationer af fænomener skaber forundring, som er en inspiration til opstilling af teorier. Statistik er en vej til at vise, om teorierne er valide eller ej. Statistik eller en anden metode er således et bindeled eller en integreret del af en analytisk arbejdsproces.

Astronomen Tycho Brahe (1546–1601) var vel en af de første, der indså denne sammenhæng. I 1572 observerede han en ny og meget stærk stjerne på himlen (det, som nu kaldes en Supernova). På den tid havde det været læren, at himlen var en konstant og uforanderlig størrelse. Observationen ledte Tycho Brahe til den overvejelse, at himlen ændres over tid. For at verificere denne teori, var det nødvendigt at foretage systematiske observationer nat efter nat af bevægelser på himlen. Dette ledte til statistikker over bevægelserne på himlen, som kunne anvendes til videre analyser og opstilling af hypoteser. Det vil sige en direkte anskueliggørelse af den postulerede sammenhæng i figuren ovenfor.

For at man kan anvende *statistisk materiale* på denne måde, må man kunne gøre følgende:

1. Teorien må kunne *operationaliseres*. Det vil sige, at den kan omsættes til målbare data
2. Det må være muligt at finde eller indsamle statistisk materiale, som reflekterer teorien

Det er vigtigt at gøre sig klart, at der i samspillet mellem teori og statistik, tillige er et samspil mellem to forskellige sæt af hypoteser, der særskilt skal undersøges. Nemlig de teoretiske og de statistiske.

2. Undervisningen i statistik før og nu

Det er således ganske store mængder af viden, som studenterne skal have styr på i forbindelse med undervisningen i statistik på økonomiuddannelserne. I det følgende foretages en subjektiv tilgang idet der drages paralleller mellem undervisningen på cand-oecon studiet i perioden fra 1980 til 1982, og forfatterens erfaringer dels som mangeårig seminarleder i beskrivende økonomi på de erhvervsøkonomiske uddannelser dels som fagansvarlig for undervisningen i statistik på de erhvervsøkonomiske uddannelser på Syddansk Universitet.

Nedenstående tabel søger at give en sammenligning mellem de to undervisningsforløb i statistik. Tilbage i begyndelsen af 1980erne var der tale om et toårigt forløb i statistik. Der var 3 ugentlige forelæsninger og 2 øvelsestimer i 4 semestre. Det det første kursus indgik imidlertid beskrivende økonomi svarende til i underkanten af 1 semester. Det vil sige, at belastningen var på ca. 3 semestre med 3 forelæsninger og 2 øvelser i 15 uger. Nogle steder eksempelvis på det daværende Handelshøjskolen i Aarhus var metodedelen omkring skrivning af opgaver udskilt i et særskilt kursus i beskrivende økonomi. Her anvendtes ca. et halvt semester på disse aktiviteter, hvorefter studenterne skrev en opgave. Denne tilgang anvendes i dag såvel på Aarhus som Syddansk Universitet.

I 2019 på Syddansk Universitet omhandler kurset "statistik" placeret på 2. semester alene den rene empiriske statistik. Det vil sige, at den statistiske metode og til dels den statistiske induktion er placeret andetsteds. Dette kursus er på 14 uger og omfatter 2 forelæsninger og 2 øvelser per uge. Dertil skal tillægges undervisning i metode. Denne udgør i et typisk kursus med en samfundsbeskrivende aktivitet ca. 15 forelæsninger og 15 øvelser. Samlet er den statistiske metodiske tilgang således på 45 forelæsninger og 45 øvelser og har en vægt på ca. 10 ECTS (heraf 5 ECTS i den rene statistikkursus). Dertil kommer et senere kursus i spørgeskemateknik som der her ses bort fra, da dette kursus tidligere lå på overbygningsuddannelsen.

Den samlede mængde konfrontationsundervisning var således i begyndelsen af 1980erne på 135 forelæsninger og 90 øvelsestimer. Det vil sige, at der i begyndelse af 1980erne var tre gange så mange forelæsninger og dobbelt så mange øvelser som det er tilfældet i dag. Vægtningen svarende til 4/20, men det var af første del, som var på 2 år

uden en større opgave. På 2 år erhverver man i dag 120 ECTS, så vægtningen i statistik svarer til 24 ECTS i dagens system. Samlet må det vurderes at disciplinen statistik tidligere fyldte mere end det er tilfældet i dag. Dette rejser en række spørgsmål. For det første om den større vægt af statistik havde sin berettigelse. For det andet om statistik i dette omfang var relevant og endelig om disciplinen statistik i dag fylder for lidt.

2.1. Emnekredse i statistik

For at undersøge det første spørgsmål ses der på en sammenligning af emnekredsene som er forsøgt opstillet i nedenstående tabel

Emnekredse 1980-1982 cand.oecon studiet, AaU		**Emnekredse 2020 HA-studiet, SDU**	
Befolkningsstatistik	Lexis skema med mere	*Ikke med*	
Statistisk metode	Problemformuleringen Planlægning indsamling Bearbejdelse (tabel og graf)	Statistisk metode	Problemformuleringen Brug af data
Statistisk induktion	Design, fordelinger Position og spredning Standardberegning Indekstal Projektioner Tidsrækkeanalyse	Deskriptiv statistik	Position og spredning Ekstremer Histogram og tidsplot
Sandsynlighedsregning		*Ikke med*	
Forventede værdier af stokastiske variabler		(skulle være i finansiering)!	
Fordelinger	Binomialfordeling Hypergeometrisk fordeling Multinominal fordeling Poissonfordeling Normalfordelingen Eksponentialfordelingen	Fordelinger	Binomialfordeling Poissonfordeling Normalfordelingen Eksponentialfordelingen
Stikprøver	Centrale grænseværdi Stikprøvens størrelse Repræsentative stikprøver Kvotientskøn	Stikprøver	Centrale grænseværdi
Hypotesetest (middelværdi, varians og andele)	Simpel en- og to-sidet 2 stikprøver En-sidet variansanalyse Goodness-of-fit Test for uafhængighed Ikke-parametriske metoder	Hypotesetest (middelværdi, varians og andele)	Simpel en- og to-sidet 2 stikprøver En-sidet variansanalyse Test for uafhængighed
Industriel statistik	Proceskontrol og -studier	*Ikke med*	
Regression	Simpel - beregning Multipel - beregning	Regression	Simpel - i regneark Multipel - i regneark Modelkontrol Modelselektion Transformation

I det gamle kursus er der en del flere emnekredse. Der er udgået flere temaer. Befolkningsstatistikken og tidsrækkerne vil næppe blive savnet i dag. Man kan nok diskutere relevansen af de ikke-parametriske metoder samt de to fordelinger, der ikke findes i dag. Endelig er industriel statistik nok heller ikke så relevant i dagens kontekst hvor store datasæt er normen.

Relevant er dog tabel- og grafteknik. Dette tema anvendes der mere end 100 sider på i Jeppesen (1972), og i Jeppesen (1975) bruges der mange sider på relationstal, vækstkvotienter, standardberegninger og lignende. Et problem i den nutidige undervisning er den meget ukritiske tilgang til udarbejdelsen af tabeller og figurer, hvor det i høj grad bliver tabuleringen fra det statistiske bureau, der bestemmer udformningen og ikke den relevante problemformulering.

I nutidens kursus i statistik er emnekredsene færre og gennemgangen nok mere overfladisk. Der er imidlertid også en væsentlig nyskabelse; nemlig den integrerede brug af regneprogrammel i undervisningen. Der er mange forskellige tilgange. På kurset ved Syddansk Universitet bruges tilføjelsesprogrammet til Excel kaldet "Dataanalyse". Da programmet er relateret til et regneark, er der selvfølgelig begrænsninger, men det vælges da næsten alle har adgang til Office-pakken fra Microsoft. Det giver mening, da langt de fleste studenter på et senere tidspunkt vil være i stand til selv at kunne foretage små statistiske undersøgelser.

Adgangen til statistik materiale har med fremkomsten af Internettet, hvor der eksempelvis er let adgang til hjemmesiderne for de nationale statistiske kontorer, også ændret sig dramatisk. Erhvervs- og samfundsbeskrivelsen har således hele ændret karakter. Den skrevne statistik, hvor man eksempelvis i statistiske efterretninger, statistisk årbog eller statistisk tiårsoversigt kunne finde færdige tabuleringer har tilsvarende været på retur. I 1980 havde sidstnævnte publikation et oplag på 50.000, mens det i 2018 var sunket til 2.500 og vel på et tidspunkt helt går digitalt.

Tilsvarende er det blevet meget lettere at fremskaffe international statistik fra eksempelvis IMF, Verdensbanken, den Europæiske Union og tilsvarende. Det er således blevet lettere at analysere statiske materialer. Endvidere er den tilgængelig lager- og proceskapacitet mangedoblet. Dette har betydet at empiriske studier baseret på behandling af store talsæt har set dagens lys eksempelvis af Thomas Piketty, Richard Florida, Steven Szymanski og Hans Rosling for bare at nævne et fåtal. Fælles for disse studier, er brugen af megen statistik, men de anvendte metoder er alene af deskriptiv natur.

Der er imidlertid også en bagside ved den lette tilgang til statistisk materiale. Først det først skal man finde og selv strukturere materialet i tabularisk og dernæst grafisk form. Dette kan gøres direkte på mange af hjemmesiderne. Typisk vil man stå med den rå tabel og dernæst skulle overveje en hensigtsmæssig bearbejdning i forhold til den givne

problemstilling. Her vil enten det statistiske bureaus hjemmeside eller et regneark typisk kunne gøre arbejdet. Problemet er, at denne bearbejdning ikke nødvendigvis er konsistent med den givne problemstilling. Det vil sige, at man på denne måde ikke nødvendigvis afkoder der svar på problemstillinger, som man søger. Dette peger på at fremtiden måske bliver mere analytisk simpel, men med anvendelse af stadig større mængder af information.

2.2. Opgaveløsningen dengang[1]

For at kunne besvare det andet spørgsmål tages der udgangspunkt i en eksamensopgave i regressionsanalyse stillet i kurset statistik II på cand.oecon studien 1. juni 1982. Opgave var formuleret som følger:

Nedenstående materiale stammer fra en bog om statistisk analyse af stikprøvedata. Følgende er givet:

$$SAK_1 = \sum_{i=1}^{25}(x_{1i} - \bar{x}_1)^2 = 8.686016 \qquad SAP_{12} = \sum_{i=1}^{25}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = 19.10464$$

$$SAK_2 = \sum_{i=1}^{25}(x_{2i} - \bar{x}_2)^2 = 102.6710 \qquad SAP_{13} = \sum_{i=1}^{25}(x_{1i} - \bar{x}_1)(x_{3i} - \bar{x}_3) = 186.764$$

$$SAK_3 = \sum_{i=1}^{25}(x_{3i} - \bar{x}_3)^2 = 8782 \qquad SAP_{23} = \sum_{i=1}^{25}(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3) = 434.770$$

$$\bar{x}_1 = \sum_{i=1}^{25} x_{i1} = 1.4044 \qquad \bar{x}_2 = \sum_{i=1}^{25} x_{2i} = 6.9306 \qquad \bar{x}_3 = \sum_{i=1}^{25} x_{3i} = 53.6$$

Lad Y = X_3 og en regressionsmodel er da givet som $X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Dvs. at boligens areal i square years afhænger af boligudgiften og indkomsten begge opgivet i 1000 USD.

- Udfør en multipel regressionsanalyse af modellen. Beregn de partielle regressionskoefficienter β_1, β_2 og konstantleddet, variansanalyse, korrelationskoefficient og signifikanstest på denne samt standardafvigelser på regressionskoefficienterne

[1] Notationen i det følgende kan være mangelfuld med hensyn til konsistens. Forfatteren bruger såvel sin egen notation som den, der anvendtes i kurset. Den daværende underviser oplyste dengang, at konsistent notation inden for disciplinen statiktik var en umulig. Dette er hermed bragt videre.

Data er givet som:

Boligens areal (sqy)	Boligudgift (1000 USD)	Indkomst (1000 USD)
Y	X1	X2
60	2.090	9.700
32	0.630	7.000
82	0.870	5.210
42	1.110	6.840
85	1.470	7.840
60	1.700	5.060
66	1.330	6.260
40	1.560	5.750
42	1.400	5.550
42	0.890	8.680
68	1.640	6.670
54	1.040	4.870
47	0.750	4.530
61	1.770	9.110
45	0.950	5.940
86	2.870	9.110
75	2.230	13.340
20	0.490	5.130
64	1.960	8.990
34	0.760	4.320
79	2.090	7.230
30	1.470	6.050
24	0.660	5.210
44	1.290	7.990
58	2.090	7.020

Note: Materialet er tilgængeligt som Excel-fil ved henvendelse til forfatteren.

Systemet på matrixform løses for vektoren $\hat{\beta}$ som:

$$Y = \hat{\beta}X \Leftrightarrow \hat{\beta} = (X'X)^{-1}Y = A^{-1}B$$

Hvor X er en matrix af X-variabler samt en vektor af éttaller til konstantleddet, mens Y er en vektor af de elementer, der skal forklares. Det er i dette tilfælde X_3 .

Inklusive konstantleddet er der 3 koefficienter der skal beregnes. Det vil sige at der skal opstilles et 3x3 system. Den mest hensigtsmæssige metode er at regne i afvigelser fra middeltallet. Det vil sige i diagonalen i afvigelser af kvadratsummer og udenfor i afvigelser af produktsummer. Det lægges der også op til, da der til opgaven er angivet disse summer for de 3 talserier.

Systemet er:

$$\begin{bmatrix} n & 0 & 0 \\ 0 & SAK_1 & SAP_{12} \\ 0 & SAP_{12} & SAK_2 \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} SAP_{33} \\ SAP_{13} \\ SAP_{23} \end{bmatrix}$$

Det vil sige, at man kan reducere til et 2x2 system, der er meget lettere at løse. Ser man ikke straks denne mulighed i løsningen er man i alvorlige problemer, og skal bruge alle data for at kunne løse opgaven! Dette er endvidere såvel tidskrævende som komplekst.

Det reducerede 2x2 har følgende udseende:

$$\begin{bmatrix} SAK_1 & SAP_{12} \\ SAP_{12} & SAK_2 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} SAP_{13} \\ SAP_{23} \end{bmatrix} \Rightarrow \begin{bmatrix} 8.686016 & 19.10464 \\ 19.10464 & 102.6710 \end{bmatrix} \times \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 186.764 \\ 434.770 \end{bmatrix}$$

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 8.686016 & 19.10464 \\ 19.10464 & 102.6710 \end{bmatrix}^{-1} \times \begin{bmatrix} 186.764 \\ 434.770 \end{bmatrix} \Leftrightarrow \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = A^{-1}B$$

Løsningen til dette system finder man så i lærebogen til kurset i matematik. Man får index koefficientmatrix benævnes A:

$$|A| = 8.686016 \times 102.6710 - 19.10464^2 = 526.81468$$

$$adjA = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} = \begin{bmatrix} 102.6710 & -19.10464 \\ -19.10464 & 8.686016 \end{bmatrix}$$

Opstilling inverst system:

$$A^{-1} = \frac{1}{|A|} \times adjA = \frac{1}{526.81468} \begin{bmatrix} 102.6710 & -19.10464 \\ -19.10464 & 8.686016 \end{bmatrix} = \begin{bmatrix} 0.1948902 & -0.0362644 \\ -0.0362644 & 0.0164878 \end{bmatrix}$$

Ved indsættelse løses for koefficienterne:

$$\begin{bmatrix} 0.1948902 & -0.0362644 \\ -0.0362644 & 0.0164878 \end{bmatrix} \times \begin{bmatrix} 186.764 \\ 434.770 \end{bmatrix} = \begin{bmatrix} 20.6318 \\ 0.3955 \end{bmatrix}$$

Da koefficientestimaterne nu er kendte og middelværdierne er givet, kan konstantleddet beregnes som:

$$\beta_0 = \overline{x}_3 - \beta_1 \overline{x}_1 - \beta_2 \overline{x}_2 = 53.6 - 20.6318 \times 1.4044 - 0.3955 \times 6.9306 = 21.88$$

Modellen får da følgende udseende:

$$Y_i = X_{3i} = 21.88 + 20.63 X_{1i} + 0.40 X_{2i} + \varepsilon_i$$

Variansanalysen er et generelt test for signifikansen af regressionskoefficienterne. Den totale effekt er givet som $SST = SAK_3 = \sum_{i=1}^{25}(x_{3i} - \bar{x}_3)^2 = 8782$. Den lineære effekt *SSTR* eller variationen mellem grupper findes idet p=2 som:

$$SSTR = \sum_{j=1}^{p} \beta_j SAP_{yj} = (20.6318 \times 186.764) + (0.3955 \times 434.770) = 4025.229$$

Da den totale variation er givet ved SAK_3 ANOVA-tabellen ser ud som:

Variation	*Kvadrat sum SS*	*Friheds-grader (fg)*	*Middelkva-dratsum (MS)*	*F-værdi*
Mellem grupper	*SSTR*= 4025.23	2	2012.61	9.31
Indenfor grupper	SSE= 4756.77	25−(2+1)=22	216.22	
Total	*SST*= 8782.00	25−1=24		

Den kritiske F-værdi kan findes i relevant tabelværk som $F_{0.05}(2,22) = 3.44 < 9.31$. H_0 hypotesen, der siger at $\beta_1 = \beta_2 = 0$ forkastes og regression giver mening.

Det er nu muligt at beregne den aggregerede korrelation (benævnt multipel R i Excel udskrifter) samt determinationskoefficienten R^2.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{4756.77}{8782.00} = 0.4584 \qquad r = \sqrt{0.4584} = 0.6771$$

Den justerede determinationskoefficient kan beregnes, men det spørges der ikke til i opgaven.

Signifikanstest på korrelationen kan undersøges ved indledningsvis at opstille hypoteserne:

H_0: $\rho = 0$ (ingen sammenhæng) $\qquad H_1$: $\rho \neq 0$ (positiv eller negativ sammenhæng)

Testeren er t-fordelt med $fg = n-2$.

$$t_{(n-2)} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.6771}{\sqrt{(1-0.6771^2)/(25-2)}} = \frac{0.6771}{0.1534} = 4.41$$

Da $t_{0.025}(23) = 2.08$ forkastes H_0 og regressionen giver mening (det skulle man også forvente fra ANOVA-tabellen).

Nu foretages begning af standardafvigelser på koefficientestimaterne. Her skal man huske at man netop i ANOVA-tabellen har fundet et estimat på den samlede standardfejl givet som $s^2 = 216.22$:

$$Var(\beta_i) = A^{-1}s^2 = \begin{bmatrix} 0.1948902 & -0.0362664 \\ -0.0362664 & 0.0164878 \end{bmatrix} \times 216.22 = \begin{bmatrix} 42.1386 \\ 3.5649 \end{bmatrix} \Rightarrow SDV(\beta_i) = \begin{bmatrix} 6.4914 \\ 1.8881 \end{bmatrix}$$

Umiddelbart ser det ud til, at den første koefficient er signifikant, mens dette ikke er gældende for den anden koefficient.

Det partielle t-test undersøger hypoteserne H_1: $\beta_i = 0$ og H_1: $\beta_i \neq 0$ ved testeren

$t = \frac{\beta_i}{SDV(\beta_i)}$ og fg = n −(p+1).

$$t_1 = \frac{20.6318}{6.4914} = 3.18 \qquad t_2 = \frac{0.3955}{1.8881} = 0.21 \qquad \text{fg=25−2=23}$$

$$t_{0.025}(23) = 2.08$$

Det ses at den første variabel er signifikant, mens dette ikke er tilfældet for den sidste variabel.

(opgaven fortsætter her med en andel del hvor man skal analysere om nogle yderligere oplysninger om husstandens størrelse øver indflydelse på resultaterne. Hvis man indarbejder disse oplysninger, kan beregningerne gentages - dette er dog undlagt for ikke at trætte læseren).

Vurdering

Opgavens tilsnit er alene det beregningstekniske, hvor man er helt tabt, hvis man ikke har husket at medbringe lærebogen i matematik til eksamen. Der er således ikke en relatering til erhvervs- eller nationaløkonomi.

Der er i eksamensteksten ikke angivet en vægtning men det vurderes at ovennævnte beregninger skulle kunne udføres på ca. 1 time.

Til et møde i Sandbjergkredsen for år tilbage berettede professor Ellen Andersen fra København Universitet engang, at man som øvet på en god dag kunne beregne hele regressionen på omkring 30 minutter. Som mindre øvet student ville denne tid nok være noget længere.

Resultaterne i opgaven kan i dag hurtigt verificeres ved anvendelse af eksempelvis Excel. Derved fremkommer følgende udskrift. Endvidere er der beregnet en korrelationsmatrix og vist residualdiagrammer.

Regressionsstatistik	
Multipel R	0.68
R-kvadreret	0.46
Justeret R-kvadreret	0.41
Standardfejl	14.70
Observationer	25

ANOVA

	fg	*SK*	*MK*	*F*	*P-værdi*
Regression	2	4025.23	2012.61	9.31	0.00
Residual	22	4756.77	216.22		
I alt	24	8782.00			

	Koefficienter	*Standardfejl*	*t-stat*	*P-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	21.88	10.51	2.08	0.05	0.08	43.68
X1 (Boligudgift)	20.63	6.49	3.18	0.00	7.17	34.09
X2 (Indkomst)	0.40	1.89	0.21	0.84	-3.52	4.31

Det kan verificeres, at alle beregningerne ovenfor er korrekte. Selve operationstiden på at udarbejde denne regression var ca. 1 minut. I tilgift får man p-værdierne, konfidensintervaller for koefficienterne og residual- samt normalfordelingsdiagram. Spørgsmålet er om detaljeret kendskab til alle beregninger er nødvendige i dag?

I perspektiv kan man se, at der ikke overraskende er en positiv indflydelse fra de 2 X-variabler. Et interessant spørgsmål er korrelationen mellem variablerne. En korrelationsmatrix kan findes som:

	Y	*X1*	*X2*
Y (størrelse)	1.00		
X1 (boligudgift)	0.68	1.00	
X2 (Indkomsten)	0.46	0.64	1.00

Det ses at boligudgiften har den største korrelation med størrelsen. Desuden er der en betydelig korrelation mellem de 2 forklarende variabler dvs. multikollinaritet. Indkomsten bør på denne baggrund undlades.

2.3. Statistiske tilgange i 2020 - hvordan vil man gøre nu?

For at kunne besvare det tredje spørgsmål tages der udgangspunkt i en eksamensopgave i regressionsanalyse stillet til en eksamen på et Negot-hold ved Syddansk Universitet juni 2019. Opgaven er i denne version udformet, som en opgave, hvor regression i regnearket anvendes som en del af løsningen. Studenterne har ca. 25 minutter til at løse denne opgave. Alle studenter har via et 365-abonnement adgang til såvel regnearket, som til tilføjelsesprogrammet "dataanalyse".

Opgaven er baseret på, at studenterne har adgang til en Excel-fil med data. Opgaven er oprindeligt stillet på engelsk, men her er den i dansk oversættelse. Opgavens tema er regionale data om sundhed, som er udviklet på Syddansk Universitet se også på adressen http://www.danskernessundhed.dk/.

Tilfældigt blev mange af de resultater, som man kan se i denne opgave, omtalt i en serie artikler i dagbladet Politiken januar og februar 2019. I denne fremstilling er svarene på spørgsmålene indarbejdet direkte i teksten. Opgaven indgik med 25 procent af eksamen svarende til at der var ca. 25 til 30 minutter til løsningen[2]. De 25 procent svarer til 5 points. I alt kunne man opnå 20 points til denne eksamen og der kræves 10 point for at bestå. De 5 points er fordel på et hovedspørgsmål på 2 points samt 6 mindre spørgsmål på hvert et halvt point.

Data I Excel-filen giver information til at kunne opstille og estimere en model for udgifterne til sundhed per indbygger i DDK i Danmarks kommuner for 2013; y-variablen til at blive undersøgt kaldes *HEALTHSPEND*. I regressionen er følgende 6 x-variabler inkluderet som det fremgår af listen her.

Variabel	*Beskrivelse*
HEALTHSPEND	Udgift per indbygger på sundhed eksklusive ekstern finansiering, DKK
UNEMP	Registeret ledighed, %
SHAREAID	Andel af befolkningen, der modtager helårs kontanthjælp, %
STRESS	Andel af befolkningen over 16 år med højt stressniveau, %
HEALTHYFOOD	Andel af befolkningen over 16 år med sundt kostmønster, %
NOACTIVITY	Andel af befolkningen over 16 år, som ikke følger WTO's minimumskrav for fysisk aktivitet, %
OVERWEIGHT	Andel af befolkningen over 16 år, som har moderat eller voldsomt overvægtsproblem givet ved BMI, %

A. Anvendt Excel til at estimere en model hvor HEALTHSPEND forklares af de 6 x-variable. Udarbejd en pæn præsentation af resultatet
2P

Resultatet ved anvendelse af Excel har følgende udseende:

Model Output

Regression Statistik	
Multiple R	0.47
R-kvadrat	0.22
R-kvadrat justeret	0.17
Standardfejl	240.14
Observationer	98

[2] En Excel-fil med det anvendte statistiske materiale er tilgængeligt ved henvendelse til forfatteren.

ANOVA

	fg	*SK*	*MK*	*F*	*P-værdi*
Regression	6	1476342	246057	4.27	0.00
Residual	91	5247773	57668		
Total	97	6724115			

	Koefficient	*Standardfejl*	*t-stat*	*P-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	2534.93	514.25	4.93	0.00	1513.43	3556.43
UNEMPLOY	109.16	37.64	2.90	0.00	34.40	183.93
SHAREAID	13.86	55.09	0.25	0.80	-95.57	123.29
STRESS	-34.35	11.14	-3.08	0.00	-56.48	-12.21
HEALTHYFOOD	-28.56	12.45	-2.29	0.02	-53.29	-3.84
NOACTIVITY	18.19	14.62	1.24	0.22	-10.86	47.24
OVERWEIGHT	-16.07	9.90	-1.62	0.11	-35.73	3.59

Med udgangspunkt i den udskrift, som du har lavet, besvar da følgende korte spørgsmål:

B. Hvilke af variable er signifikante ved et 5 % niveau?
½P

Variablerne UNEMPLOY, STRESS og HEALTHYFOOD har alle en p-værdi under 0.05, så disse er signifikante.

C. Hvilken hypotese undersøges i ANOVA-tabellen i regressionsoutputtet?
½P

ANOVA-tabellen giver en generelt fælles (joint) test af de inkluderede k forklarende variable. Hypoteserne kan opskrives som:

H_0: $\beta_{UNEMP} = \beta_{SHAREAID} = \beta_{STRESS} = \beta_{HEALTHYFOOD} = \beta_{NOACTIVITY} = \beta_{OVERWEIGHT} = 0$
H_1: Minimum en variabel er forskellig fra nul

Som det fremgår af output, så forkastes H_0 (variables listet under spørgsmål i B er signifikant forskellige fra nul).

D. Hvad er den justerede R-kvadrat lig med og hvordan adskiller den sig fra R-kvadrat?
½P

Den justerede R-kvadrat er lig med 0.17. Dette betyder at x-variablerne forklarer 17 procent af variationen i y når der er korrigeret af antallet af frihedsgrader. På denne måde

opnås et mere retvisende billede af den forklarende variation. Den justerede R-kvadrat er altid mindre end den ikke justerede R-kvadrat.

E. Standardfejlen af nærværende model er lig med 240.14. En alnativ model er blevet estimeret med en standardfejl der er lig med 362.11. Skal denne model anvendes I stedet for den nuværende?

½P

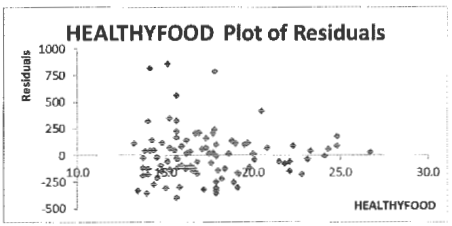
Standardfejlen skal være så lav som muligt. Da 240.14 < 362.11 er nærværende model bedre end den alternative.

F. Arbejdsløshed resulterer en forøgelse af udgifterne til sundhedspleje per capita på 109.16 DKK. Kunne denne koefficient tænkes at være lig med 165.- DKK?

½P

For variablen UNEMPLOY er 95 % konfidensintervallet omkring koefficientestimatet lig med [34.40;183.93] En værdi på 165,- DKK findes indenfor variationen givet med dette interval. Det vil sige, at koefficienten godt kan tage denne værdi.

G. Residualdiagrammet mellem sundhedsudgifterne per capita *(HEALTHSPEND)* og andelen af befolkningen med gode kostvaner *(HEALTYFOOD)* har følgende udseende:



Ligner residualerne hvid støj? **½P**

Residualdiagrammet undersøger validiteten af antagelsen om, at de medgåede data i regressionen er normalfordelte og uafhængige. Er dette tilfældet vil residualerne udvise hvid støj. Det vil sige at de er uafhængige og normalfordelte med middelværdi lig nul og konstant varians

Det fremgår, at dette ikke er tilfældet i residualdiagrammet ovenfor. Data virker stationære omkring middelværdien på nul, men variansen er ikke konstant og faldende med stigende værdi af *HEALTHYFOOD*. Dette kaldes også *heteroscedasticitet.*

En anden tilgang til at stille opgaver findes ved HA-studiet ved Syddansk Universitet. Her bruges en 48-timers hjemmeopgave baseret på et lignende datasæt. Her skal studenterne selvfølgelig præstere anderles og besvare spørgsmål inden for mange aspekter af hypotese tests, krydsfordelinger med mere og arbejde med fordelinger for uden at kunne modelselektere en model, der ligner den ovenstående. Endelig skal studenterne kunne formulere et resumé, hvor de på en letlæselig måde sammenfatter resultaterne af, hvad de har fundet[3].

3. Afslutning og udblik

Kurser i statistik på de samfundsøkonomiske uddannelser generelt er reduceret ganske betragteligt i omfang såvel med hensyn til forelæsninger og øvelser samt tematisk. Endvidere har en større kompression af stoffet medført at den dybere forståelse af de anvendte metodiske tilgange er reduceret.

I samme tidsrum er mængden af tilgængeligt statistik materiale, der er det applikative udgangspunkt for undervisningen i statistik, forøget voldsomt. Fokus i undervisningen rykket mod det mere analytiske og mindre mod det regnetekniske i forhold til tidligere. Her var vægten på at forstå og udføre relevante beregninger. Nu udføres de fleste af disse ved brug af computer især regnearksbaseret, som det fremgår af nuværende opgaver på HA-og Negot-uddannelserne ved Syddansk Universitet.

Denne udvikling er et resultat af dels bedre informationsteknologi dels færre ressourcer til emnekredse relateret til statistik. Set i lyset af studenternes kundskaber i matematik, hvor omkring 75 procent har adgangsgivende eksamen på matematik B niveauet, virker prioriteringen hensigtsmæssig. Dækningen af temaer som fordelinger, hypotesetests og regression vurderes således at være nogenlunde tilfredsstillende. Derimod er det ikke tilfredsstillende, at studenterne ikke har kendskab til ordentligt at udtage stikprøver fra totalpopulationer. Dette også set i lyset af, at de studerende i løbende bombarderes med diverse spørgeskemaer relateret til evalueringen af undervisningen.

Studenternes tilgang til den statistiske metode og induktion er derimod mangelfuld. Tilgangen til statistiske tabuleringer og grafiske virkemidler er mangelfulde og ligger normalt uden for rammerne af et kursus i statistik. Ligeledes er studenternes kendskab til indekstal, standardberegning, vækstkvotienter og lignende mangelfuldt. Disse temaer burde føres tilbage kurset i statistik, der bør tildeles flere ressourcer såvel i form af forelæsninger, men især i form af øvelser.

[3] Dette er alene meget summarisk forklaring. Detaljerede opgaver fremsendes gerne ved henvendelse til forfatteren. Foruden forhold relateret til sundhed her temaer for opgaver blandt andet været den europæiske banksektor, udbredelse af solceller i Danmark, udenlandske investeringers betydning for økonomisk vækst, innovation og konvergens i Europa danske eksportvirksomheder og udbredelsen af skyggeøkonomi.

Hvad blev tabt på vejen? Den tekniske side af befolkningsstatistikken, den industrielle statistik og de elementære tidsrækker vil nok ikke blive savnet af underviserne og studenterne.

Fremtidens kurser i statistik bør tage hånd om en bedre analyse og mere kritisk med de store og lettilgængelige mængder af statistik på Internettet. Desuden bør studenterne blive bedre til dels at tage stikprøver dels at kunne præsenteret i statistisk materiale, så det er i god harmoni med en given problemformulering.

Litteratur

Barrow, M. (2017): "Statistics for Economics, Accounting and Business Studies". 7th edition. Pearson.

Bowerman, B.L., R. T. O'Connell & E. S. Murphree

Bowerman, B.L., R. T. O'Connell, E. S. Murphree & J. B. Orris (2015): "Essentials of Business Statistics". 5th edition. McGraw-Hill.

Bryman, A. & E. Bell (2015): "Business Research Methods". 4th edition. Oxford University Press.

Florida, R. (2017): "The New Urban Crisis". Basic Books.

Erik Harsaae (1981): "Industriel statistik og repræsentative undersøgelser". Statistisk Institut, Aarhus Universitet.

Jeppesen J. (1972): "Statistisk metode". 2. udgave. Akademisk forlag.

Jeppesen, J. (1975): "Statistisk Induktion". 2. udgave. Akademisk forlag.

Linderoth, L. & J. Bentzen (2009): "Udarbejdelse af rapporter i beskrivende økonomi". Handelsvidenskab bogforlaget.

Piketty, T. (2015): "Capital in the Twnenty-First Century". Harvard University Press.

Rosling, H., O. Rosling & A. R. Rönnlund (2018): "Factfulness". Sceptre.

Studiehåndbog Økonomi 1980 og 1981, Økonomisk Institut Aarhus Universitet.

Syddansk Universitet (2019): "Fagbeskrivelse for kurset "matematik og statistik"". Se https://odin.sdu.dk/sitecore/index.php?a=fagbesk&id=58316&listid=4925&lang=da

Szymanski, S. (2015): "Money and Football - A Soccernomics Guide". Nation Books.

Sørensen, N.K., (2019): "Undersøgelse af karakterer i matematik og mikroøkonomi på HA/BA-uddannelserne ved Syddansk Universitet". I Linde, P. (editor): "Symposium for anvendt statistik 2019", side 207–220. SAS Institute, ISBN 978-87-7904-359-6.

References

Alstadsæter, A., Johannesen, N., and Zucman, G. (2017). 'Tax evasion and inequality'. Working Paper 23772, National Bureau of Economic Research.

Atkinson, A. B. and Bourguignon, F. (2015). 'Introduction: Income distribution today'. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2, pages xvii – lxiv. Elsevier.

Causa, O., Hermansen, M., Ruiz, N., Klein, C., and Smidova, Z. (2016). 'Inequality in Denmark through the looking glass'. Working Paper 1341.

Esping-Andersen, G. (2017). 'Kapitel 8 - Multidimensional omfordeling i velfærdsstater'. In Ploug, N., editor, *Økonomisk ulighed i Danmark*, pages 147–164. Jurist- og Økonomiforbundets Forlag.

Jäntti, M. and Jenkins, S. P. (2015). 'Chapter 10 - Income Mobility'. In Atkinson, A. B. and Bourguignon, F., editors, *Handbook of Income Distribution*, volume 2, pages 807 – 935. Elsevier.

Kuznets, S. (1955). 'Economic growth and income inequality'. *American Economic Review*, 45(1):1.

Milanovic, B. (2016). *Global Inequality: A New Approach for the Age of Globalization*. Harvard University Press.

Piketty, T. (2014). *Kapitalen i det 21. århundrede*. Gyldendal, København, 1. edition.

Piketty, T., Saez, E., and Zucman, G. (2018). 'Distributional National Accounts: Methods and Estimates for the United States*'. *The Quarterly Journal of Economics*, 133(2):553–609.

Quitzau, J. (2017). 'Kapitel 2 - Indkomstfordelingen i Danmark'. In Ploug, N., editor, *Økonomisk ulighed i Danmark*, pages 33–61. Jurist- og Økonomiforbundets Forlag.

Stiglitz, J. E. (2013). *The price of inequality*. W.W. Norton & Co, 1 edition.

Økonomiske Råd, D. (2016). *Dansk økonomi efteråret 2016*. De Økonomiske Råd.

Repræsentativitet i paneler med løbende udskiftning

- En løsning, der virker universelt, når populationen ikke er statisk

Peter Linde, Det Nationale Forskningscenter for Arbejdsmiljø

Baggrund

Med simpel tilfældig udvælgelse i hele populationen får man en universel repræsentativ stikprøve. Man kan fx i en tværsnitsstikprøve have udvalgt 1.000 personer i alderen 18-65 år simpelt tilfældigt fra CPR. Året efter ønsker man at lave en tilsvarende repræsentativ tværsnitsundersøgelse, og for at kunne analysere ændringer på personniveau, vil man danna et panel og genkontakte en del af stikprøven fra forrige år, der er i live, bosiddende i Danmark og nu er 19-65 år. Måske er stikprøven stratificeret, så man har udvalgt ekstra mange i nogle geografiske områder, i særlige jobgrupper eller blandt unge. Hvordan udvælger man stikprøven som året før uden af få skævheder i den nye tværsnitsstikprøve? Og hvordan vægter man for det design, man er endt med? Det er normalt håndterbart den første gang, man skal trække en ny justeret stikprøve, der fortsat kan bruges repræsentativ som tværsnitsundersøgelse. I eksemplet ovenfor skal man sikre, der er nok 18-årige i den nye stikprøve, da alle er blevet ét år ældre. Der er også nye i populationen blandt de 19-65-årige, der er indvandret eller kommet tilbage til Danmark. Hver gang man skal gentage øvelsen, bliver det dog mere og mere komplekst, hvis man ikke har en systematisk løsning. Særligt hvis der også er strata, hvor man har trukket ekstra mange, fx i nogle geografiske områder, og nogle flytter fra og til disse områder (strata).

Hvorfor genkontakte overhovedet?

Der er en række gode grunde til at genkontakte en del af de udvalgte, hvis undersøgelsen har samme spørgsmål og population, som man gentager flere gange. De vigtigste grunde er:

- Man kan analysere ændringer på micro-niveau (personniveau).
- Sammenligningen mellem to tværsnit får mindre varians, fordi en del af stikprøverne er parvise sammenligninger af de samme personer. Hvis fx halvdelen er genudvalgt, kan variansen blive ned til 50% af det den ellers ville have været. Normalt kun den halve gevinst, men det er bestemt også værd at tage med.
- Omkostningerne ved dataindsamlingen er mindre, fordi man kan genbruge metainformation, om hvordan fik kontakt og svar. Fx telefonnummer, træffetidspunkt eller e-mail.

- Opnåelsen stiger, hvis det gøres rigtigt. Det kan måske undre, men en del af dem, der ikke svarede sidste gang, svarer anden gang de bliver spurgt. De, der svarede sidste gang, svarer i høj grad også næste gang, selv om man mister nogle. Den samlede effekt er normalt positiv – ikke mindst hvis man skriver to forskellige breve til de to grupper.

Der er mange eksempler på ovenstående positive effekter af genudvælgelse. Her kan fx nævnes arrangementet i Selskab for Surveyforskning den 19. september 2018 – www.surveyselskab.dk. Kilde 1.

Der er også en risiko ved at genudvælge. Den ene er, hvis man ikke trækker stikprøven rigtigt. Det giver denne artikel en løsning på. Den anden risiko er, hvis man genkontakte alle man kan. Det kan i første omgang virke effektivt, men på et tidspunkt blive de fleste trætte af at blive kontaktet. Omkring fire gange er normalt det bedste, og så skal man gå i gang allerede den første gang med at frigive en ¼-del, fordi ellers har man ikke en løsning, når der er gået fire gange. Det er der også en løsning på i denne artikel. Endelig skal man genkontakte både de personer, der svarede og de, der ikke svarede. Man få en ekstra bias, hvis man kun kontakter de, der svarede sidst på grund af 'renters rente' i bortfaldet. Det bliver for hver gang, man gentager, mere og mere fx de højst uddannede, danskerne og kvinderne, der deltager.

To eksempler, hvor det kan opstå udfordringer

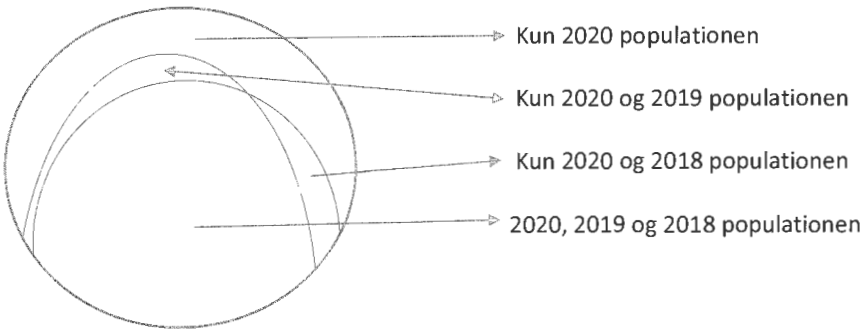
Først kigger vi på udfordringen i et meget almindeligt eksempel, hvor flere stikprøver blandes. Vi har en population fra 2020 og har trukket en simpel tilfældig stikprøve A, samt en population fra 2019, hvor man dengang trak en simpel tilfældig stikprøve B, og endelig en population fra 2018, hvor man dengang trak en simpel tilfældig stikprøve C.

I 2020 kontaktes de fra stikprøven A, der er udvalgt direkte fra 2020 populationen, den del af 2019 stikprøven B, der findes i 2020 populationen (denne stikprøve kalder vi B'), og den del af 2018 stikprøven C, der findes i 2020 populationen (denne stikprøve kalder vi C'). Samlet kontakter vi således A, B' og C', som alle findes i 2020 populationen.

I Figur 1 kan man se, at 2020 populationen kan opdeltes i fire strata. De, der kun tilhører 2020 populationen, de der tilhører både 2020, 2019 og 2018 populationen, de der kun tilhører 2020 og 2019 populationen og de der kun tilhører 2020 og 2018 populationen. Udvalgssandsynligheden vil være meget større for dem, der tilhører fællesmængden af de tre populationer, for man her har mulighed for at blive valgt

tre gange gennem både stikprøve A, B' og C'. Udvalgssandsynligheden vil være mindst for dem, der kun tilhører 2020 populationen, fordi man kun vælges gennem stikprøve A. Man har derfor fået en stikprøve med fire strata og i hvert stratum skal der vægtes med forholdet mellem populationen N i stratummet og stikprøven n i stratummet. Vægten er de fire forhold N/n i stratummet og det giver fire forskellige vægte. Hvis man ikke vægter, vil der komme en bias, og tværsnittet kan hverken bruges til at sige noget retvisende om 2020 populationen eller sammenlignes med tidligere tværsnitsundersøgelser.

Figur 1. Populationen i 2020 opdelt efter om man også tilhørte populationerne i 2019 og 2018



En måde at løse ovenstående udfordring kunne være i 2020 at supplere/justere stratum for stratum, så der er den samme udvalgssandsynlighed. Det vender vi tilbage til. Først et eksempel på udfordringer med at supplere en stikprøve.

Første år er en simpel tilfældig udvælgelse i hele populationen. Populationen kan være de 18-70-årige beskæftigede eller en anden gruppe. Populationen er på 1.000.000 og vi ønsker at vælge 0,1%, dvs. 1.000 personer. Det går fint det første år. Hvis alle svarer, har vi vægten 1.000 bag hvert svar. Næste gang/år er populationen ændret lidt – den er ikke statisk. 1.050.000 tilhører nu populationen. Vi har fået lidt ekstra midler og vil gerne vælge 1.100 denne gang. Af de 1.000.000 fra året

før er 50.000 ikke mere i populationen, så 950.000 er fortsat i populationen og 100.000 kommet til. Der er i alt 1.050.000 i den nye population. Af de 1.000 udvalgte fra sidste udvælgelse er 950 i den nye population. 50 er ikke mere aktive i populationen og kontaktes ikke.
Man kan nu foreslå løsning 1: Vi vælger de ekstra 150, der er råd til, i hele den nye population på 1.050.000 – dog ikke blandt de 950 allerede udvalgte.
Løsning 2: Vi vælger de ekstra 150 blandt de 100.000, der er kommet til population.
Løsning 3. Vi vælger først 100 ekstra blandet de 100.000, der er kommet til populationen. Herved er der balance fordi alle i stratummet får vægten 1.000 ligesom de 950, der er med igen. Til sidst vælges 50 i hele den nye population på 1.050.000. Dette bibeholder balancen.

Løsningen 1 vil samlet give få udvalgte blandt de 100.000, der er kom til populationen. De 150 vil ca. fordele sig med 14 fra de nye 100.000 og 136 fra de overlevende 950.000. Dvs. 136 + 950 = 1.086 fra de 950.000 overlevende i populationen og 14 fra de 100.000 nye. Det giver to vægte: 950.000/1.086=875 og 100.000/14=7.142. Dvs. to opregningsstrata med to meget forskellige vægte. Hvis man ikke vægter, taber man repræsentativitet. Meget forskellige vægte øger stikprøvevariansen – det vil derfor være efficient at udvælge så de bliver mere ens.
Løsning 2 vil også give to opregningsstrata, hvis repræsentativiteten skal sikres, men ikke helt så forskellige vægte. Nemlig 950.000/950=1.000 og 100.000/150=667.
Løsning 3 er **den selvvejede løsning**. Kræver ikke to opregningsstrata. Alle har samme vægt: 1.050.000/1.100=955.

Allerede med ovenstående eksempler er det klart, at der let kan opstå problemer med hensyn til repræsentativiteten, og det for hvert år bliver mere komplekst. Hvis der også er strata med forskellige udvalgssandsynligheder, kan det gå helt galt. Man kan fx have udvalgt 1.000 blandet 2.000.000 beskæftigede og 1.000 blandt 200.000 ledige. Undersøgelsen skal sammenligne ledige og beskæftigede, og samtidig sige noget om hele populationen af dem, der er til rådighed på arbejdsmarkedet. Det stratificerede design kræver vægte, da der er udvalg 1:2.000 af de beskæftigede og 1:200 af de ledige. Næste gang vil der være bevægelse blandt de to strata – populationen er ikke statisk. Nogle beskæftigede bliver ledige og vil uden ekstra stikprøve for denne gruppe have en større vægt end de andre ledige Nogle ledige kommer i beskæftigelse og vil uden nedjusteringer for denne gruppe have en mindre vægt end de andre ledige. Man må ikke give dem samme vægt, fordi de repræsenterer forskellige delgrupper – det giver bias. Medmindre man justerer stikprøven, beholder man altid sin oprindelige udtræksvægt – der er dem man repræsenterer.

De skal enten beholde deres udtræksvægt, eller også skal der vælges ekstra blandet de, der er blevet ledige, og fravælges nogle af dem, der kommet i beskæftigelse, så der igen er balance inden for strata.

Løsningen – samordnet udvælgelse: SAMU.

Sveriges Statistik udviklede i 1980`erne et universelt system til at koordinere både person- og virksomhedsstikprøver – SAMU (kilde 2). Ideen er lige så simpel, som den er effektiv. Alle personer får tilordnet et (deres) tilfældige SAMU-tal mellem 0 og 1. Hver gang man skal trække en stikprøve for en undersøgelse starter man fx i 0, forsætter med at udvælge stigende SAMU-tal, indtil når det antal, der er kvoten for, hvor mange man vil vælge. Hvis man fx skal udvælge 1.000 af 100.000, starter man i 0 og op til omkring 0,01 svarende til at 1.000 af 100.000 cirka svarer til 1% udvælges. Hvis man derudover vil have, at de udvalgte deltager op til fire gange og bagefter fritages, lægger man fx 0,1 til ¼-del af alle de personlige SAMU-tal. Hvis man derved kommer over 1, trække man 1 fra, så man igen får et tal mellem 0 og 1. På denne måde kan man sikre, at en ¼-del får et så højt SAMU-tal, at de reelt er fritaget for mange, mange år. Næste gang man gentager undersøgelsen lægger man 0,1 til en anden ¼-del af populationen. De, der forlader populationen, ligge jævnt i hele intervallet 0 til 1, så de udgår bare. Nye i populationen får et tilfældigt SAMU-tal mellem 0 og 1, der er deres.

Det eneste, der er udfordringen, er at holde styr på hvilken ¼-del, der næste gang skal have lagt 0,1 til. Det løser man ved at afrunde det tilfældigt SAMU-tal til fx 20 decimaler. På plads nummer 21 og 22 indsætter man tilfældigt en gang for alle tallene 00, 01, 02 osv. op til 99 efter tur. Første gang lægger man 0,1 til for de med 00-24, næste gang 25-49, næste gang 50-74 og endelig 75-99, indtil man kommer tilbage til 0 til 24 på position 21 og 22. Hver gang man skal lave en udvælgelse, tager man de to tal på position 21 og 22 fra, og lægger 0,1 til SAMU-tallet efter princippet ovenfor, udvælger og sætter tallet på position 21 og 22 på igen. Man skal her huske at gemme det samlede tal som tekststreng og ikke et numerisk tal, så 0 til sidst ikke forsvinder

Hvis der er strata, udvælger man inden for hvert stratum indtil den ønskede kvote. De, der har flyttet strata, bliver enten udvalgt med øget chance, hvis det nye strata har en højere udvalgssandsynlighed – eller modsat fritaget, hvis det nye har en lavere udvalgssandsynlighed. Det hele er i balance samtidig med dette selvvejede design har den størst mulige grad af gentagelse fra den ene gang til den anden.

Koordinering af stikprøver
SAMU kan ikke kun bruges til korrekt at supplere stikprøver, sikre repræsentativitet, styre fritagelsen og optimere gentagelser. SAMU kan også bruges til at koordinere stikprøver. Hvis man har to store stikprøver og ønsker, de ikke skal påvirke hinandens opnåelse negativt, sørger man for deres to startværdier er forskellige, fx 0,00 og 0,25. Hvis man ønsker, de skal kunne flettes på hinanden, sørger man for de har samme startværdi.

Den første SAMU-udvælgelse
Det er aldrig for sent at bruge SAMU, hvis man gerne vil bruge det fremadrettet. Den viden man har om sin stikprøves historie bruges til at dele op i strata – jf. eksemplerne ovenfor. Inden for hvert stratum har man en population og en aktiv stikprøve. Hvis den aktive stikprøve fx er 1% af populationen får den et tilfældigt SAMU-tal mellem 0,00 og 0,01. Den resterende del af populationen et tilfældigt tal mellem 0,01 og 1,00. Sådan kan man give SAMU-tal for hvert stratum, og man er i gang med SAMU.

Hvad hvis man ikke kender hele populationen?
Som forsker eller privat analyseenhed kan man få repræsentative tilfældige stikprøver fra CPR med cpr-numre, men man kender ikke hele populationen. Det betyder, man ikke kan give et tilfældigt SAMU-tal til hele populationen. Her kan man udnytte, at CPR-nummeret skal opfylde en modulus-11 regel, så man kan danne alle mulige CPR-numre og give dem et tilfældigt SAMU-tal, og så bruge dette på de stikprøver, man har adgang til. Det skal nævnes, at der er 18 fødselsdatoer, der siden 2007 har haft for få mulige cpr-numre, der opfylder 11-modulus reglen og reglen ikke gælder her, men dette er trods alt en mega lille del af populationen.

Kilder

1) Rullende paneler og konsekvenser for variansberegning, bortfald og opregning – eksempler fra Arbejdskraftundersøgelsen. Helene Feveile, Metode, Danmarks Statistik. Selskab for Surveyforskning den 19. september 2018 – www.surveyselskab.dk.
2) SAMU. The system for Co-ordination of Samples from Business Register at Statistics Sweden. Esbjørn Ohlson. R&D report 1992.18

A Vine Copula Panel Model For Day-Ahead Electricity Prices

Janus S. Valberg-Madsen

This is a joint work with Esben Høg, Troels S. Christensen, and Anca Picalabu.

1 Background

On the German day-ahead power market, every day around noon, hourly electricity prices for the following day are set based on bids and offers submitted to the exchange by market agents. As those 24 hourly prices are decided from the same information set, it's not appropriate to consider them as a one-dimensional, hourly time series, but rather it should be treated as a 24-dimensional, daily time series.

In the literature, so-called *base prices*, daily averages of the day-ahead prices, are commonly considered when modelling prices (see e.g. Escribano et al. (2011) or Weron (2014)). In the following, we fit a joint model of the prices modelled as 24 individual daily time series linked together with a vine copula. Using the model, we consider the payoff distribution of a forward contract that depends on prices of individual hours.

2 The model

2.1 Marginal models

When modelling stock prices, one typically considers log-returns, but this doesn't work directly for electricity prices for two reasons. Firstly, electricity is a largely unstorable commodity, and thus it does not make sense to consider the "return" incurred in buying electricity in one period and selling it in another. Secondly, electricity prices can be negative, for example when the wind production is very high.

We accomodate for this by instead considering log-prices with an offset, i.e.

$$p_{t,h} := \log\left(P_{t,h} + K\right), \tag{1}$$

for $t \in \{1,\ldots,T\}$, $h \in \{1,\ldots,24\}$, and K is chosen to reflect bid restrictions on the market. For the German power spot market, bid and ask prices must be greater than -500 EUR, so choosing $K = 500$ will ensure that the above adjusted prices are always well-defined.

The adjusted prices are then modelled for each hour as a daily time series, split into a deterministic, seasonal component and a stochastic serial component:

$$p_{t,h} = s_{t,h} + X_{t,h}, \tag{2}$$

where the seasonal component, $s_{t,h}$, is modelled as

$$s_{t,h} = a_{0,h} + a_{1,h}t + \sum_{f\in\Phi} \left(b_{1,h,f}\sin(2\pi t f) + b_{2,h,f}\cos(2\pi t f)\right) + c_{w,h}w_t, \tag{3}$$

where $\Phi = \{1/365, 2/365\}$ is a set of frequencies corresponding to annual and semiannual cycles, and w_t is a factor variable coding for weekdays. The serial component, $X_{t,h}$, is modelled as an ARMA-GARCH process, with both ARMA and GARCH orders up to (1,1), i.e.

$$X_{t,h} = \phi_h X_{t-1,h} + \theta_h \varepsilon_{t-1,h} + \varepsilon_{t,h}, \tag{4}$$

$$\varepsilon_{t,h} = \sigma_{t,h} z_{t,h}, \tag{5}$$

where $z_{t,h}$ is a stationary noise process. For the type of GARCH process, the GJR- and exponential GARCH models are considered, i.e.

$$\sigma^2_{t,h} = \omega_h + \left(\alpha_h + \gamma_h \mathbb{1}[\varepsilon_{t-1,h} \le 0]\right)\varepsilon^2_{t-1,h} + \beta_h \sigma^2_{t-1,h} \tag{6}$$

or

$$\log(\sigma^2_{t,h}) = \omega_h + \alpha_h z_{t-1,h} + \gamma_h\left(|z_{t-1,h}| - \mathbb{E}|z_{t-1,h}|\right) + \beta_h \log(\sigma^2_{t-1,h}) \tag{7}$$

and for conditional distributions, we consider the Student's t-distribution and its skewed variant. For each series, this nets a total of 48 models, which are estimated with maximum likelihood methods and chosen between with BIC.

The residuals from the marginal models are transformed into uniformly distributed variables using the estimated distribution functions and used as input for the joint model.

2.2 Joint model

The pivotal object in the model is the *copula*, which is defined as a distribution function whose marginal distributions are all uniform on the unit interval. The significance of the model comes from the below result of (Sklar, 1959):

Theorem 2.1 (Sklar's theorem). *Let $\boldsymbol{X}$ be a d-dimensional random vector with joint distribution function F and marginals $F_1, \ldots, F_d$. Then there exists a d-copula C such that $\forall \boldsymbol{x} \in \mathbb{R}^d$,*

$$F(\boldsymbol{x}) = C\big(F_1(x_1), \ldots, F_d(x_d)\big),$$

and if the marginals are all continuous, then C is uniquely defined.

The practical consequence of this result is that one may consider the marginal and joint analysis separately—first defining marginal models for each variable and then tying it all together afterwards using a copula.

This class of models can be made even more flexible by considering so-called *vine copulas*, first introduced by (Joe, 1994) and later systematised by (Bedford and Cooke, 2002), in which 2-copulas are used as building blocks in the construction of a multi-dimensional joint distribution. The full joint density can be factored into a product of marginal densities and 2-copulas, represented as a sequence of trees such that

- The nodes in the first tree are the marginal densities, and the edges are 2-copulas linking them pairwise together
- In subsequent trees, the nodes are the edges from the previous tree, and the edges are conditional 2-copulas between them
- Nodes in a tree can only be joined by an edge if the nodes share a node from the previous tree

The flexibility over the regular d-dimensional copula comes from the fact that the 2-copulas used as the building blocks in the vine need not be the same, so one can employ different families of copulas for different conditional pairs. For our purpose, we consider Gaussian, Student's t-, Clayton, Gumbel, Frank, and Joe copulas.

3 Data

The data used consists of hourly day-ahead electricity prices in EUR/MW for Germany in the interval [2015-01-06, 2018-12-31], downloaded from the European Network of Transmission System Operators for Electricity's Transparency Platform, regulation article 12.1.D (details about these articles can be found on (ENTSO-E, 2019)). Germany is part of a bidding zone together with Luxembourg, and prior to 2018-09-30, Austria was also part of that bidding zone, so to obtain historical data for German day-ahead prices, DE-AT-LU bidding zone prices have been downloaded for 2015-2018 and DE-LU bidding zone prices have been downloaded for 2018-2019.

Due to the transition to daylight savings time, hour 3 observations are missing every

year, which are interpolated from the adjecent hour 2 and 4 observations. Likewise, there are duplicate hour 3 observations on the days where daylight savings time ends, and these are averaged into a single observation. This nets 1456 daily observations of 24 hourly prices.

4 Results

The selected serial models for each hour are listed in Table 1. Notably, for most of the hours, orders (1,1) are selected for both the ARMA and GARCH component, and in all but one case, the skewed t-distribution is preferred over the symmetric one.

Hour	Model	Hour	Model
1	ARMA(1,1)-GJR-GARCH(1,1)-st	13	ARMA(1,1)-GJR-GARCH(1,0)-st
2	ARMA(1,1)-GJR-GARCH(1,1)-st	14	ARMA(1,1)-GJR-GARCH(1,0)-st
3	ARMA(1,1)-E-GARCH(1,1)-st	15	AR(1)-GJR-GARCH(1,0)-st
4	ARMA(1,1)-GJR-GARCH(1,1)-st	16	ARMA(1,1)-GJR-GARCH(1,0)-st
5	ARMA(1,1)-E-GARCH(1,1)-st	17	ARMA(1,1)-GJR-GARCH(1,1)-st
6	ARMA(1,1)-GJR-GARCH(1,1)-st	18	ARMA(1,1)-E-GARCH(1,1)-st
7	ARMA(1,1)-E-GARCH(1,1)-st	19	ARMA(1,1)-GJR-GARCH(1,1)-t
8	ARMA(1,1)-E-GARCH(1,1)-st	20	ARMA(1,1)-GJR-GARCH(1,1)-st
9	ARMA(1,1)-E-GARCH(1,1)-st	21	ARMA(1,1)-GJR-GARCH(1,1)-st
10	ARMA(1,1)-E-GARCH(1,1)-st	22	ARMA(1,1)-GJR-GARCH(1,1)-st
11	ARMA(1,1)-E-GARCH(1,1)-st	23	ARMA(1,1)-E-GARCH(1,1)-st
12	ARMA(1,1)-E-GARCH(1,1)-st	24	ARMA(1,1)-E-GARCH(1,1)-st

Table 1: *Overview of chosen serial models for each hour*

The conditional, standardised residuals are transformed into uniform variables using each respective, estimated conditional distribution function, which are used to estimate the vine copula. A summary of the copulas in the first tree of the vine is listed in Table 2.

Notably, for all but one edge, the Student's t-copula is selected, but the shape parameter range from 3.28 to 15.57. In subsequent trees, the dependence is not as strong as in the first, but the vine cannot be truncated until the fourth-to-last tree, i.e. where all copulas of the following trees are made up of only independence copulas.

Edge	Copula	τ	Edge	Copula	τ
3–2	t ($\rho = 0.96$, $\nu = 4.66$)	0.82	14–13	t ($\rho = 0.96$, $\nu = 3.65$)	0.81
3–1	t ($\rho = 0.83$, $\nu = 13.9$)	0.63	15–14	t ($\rho = 0.96$, $\nu = 3.73$)	0.83
4–3	t ($\rho = 0.97$, $\nu = 3.28$)	0.84	16–15	t ($\rho = 0.96$, $\nu = 3.57$)	0.82
5–4	t ($\rho = 0.97$, $\nu = 4.23$)	0.85	17–16	t ($\rho = 0.94$, $\nu = 5.32$)	0.78
6–5	t ($\rho = 0.94$, $\nu = 4.67$)	0.77	18–17	t ($\rho = 0.9$, $\nu = 6.2$)	0.72
7–6	t ($\rho = 0.83$, $\nu = 13.09$)	0.62	19–18	t ($\rho = 0.9$, $\nu = 6.34$)	0.71
8–7	t ($\rho = 0.9$, $\nu = 6.23$)	0.72	20–19	t ($\rho = 0.92$, $\nu = 5.92$)	0.75
9–8	Survival Gumbel ($\alpha = 4.37$)	0.77	21–20	t ($\rho = 0.93$, $\nu = 6.35$)	0.76
10–9	t ($\rho = 0.92$, $\nu = 6.32$)	0.75	22–21	t ($\rho = 0.93$, $\nu = 4.03$)	0.75
11–10	t ($\rho = 0.95$, $\nu = 5.55$)	0.79	23–22	t ($\rho = 0.92$, $\nu = 4.56$)	0.74
12–11	t ($\rho = 0.96$, $\nu = 4.15$)	0.83	24–23	t ($\rho = 0.86$, $\nu = 15.57$)	0.66
13–12	t ($\rho = 0.94$, $\nu = 14.83$)	0.79			

Table 2: *Overview of the 2-copulas selected for the first tree of the vine with estimated parameters and Kendall's τ*

4.1 Predicting payoff distributions

The setup for the simulated payoff is as follows. We enter into a contract with a counterparty to buy electricity on certain hours of certain days in the future for a fixed, agreed upon price called the *forward price*. These hours need not be contiguous, and they need not be the same from day to day. Let $F(t, T_1, T_2)$ be the forward price in EUR/MWh determined at time t for delivery in the period $[T_1, T_2]$ that has the payoff

$$\sum_{s=T_1}^{T_2} \sum_{h \in H(s)} \left(P_{s,h} - F(t, T_1, T_2) \right), \tag{8}$$

where $H(s)$ is the set of hours over which the contract is active on a given day, and $P_{s,h}$ is the spot price for a given hour of a given day. In other words, on active hours where the spot price $P_{s,h}$ is higher than the forward price $F(t, T_1, T_2)$, the position corresponds to a profit.

Using the full joint model, the spot prices are simulated for the interval [2019-01-01, 2019-07-31], netting 212 observations of 24 hours. These are then used to calculate the total payoff on the contract $F(t, T_1, T_2)$. This is repeated 500 times to simulate the distribution of the payoff. As an example, consider a forward contract written on 2018-12-31 for delivery on hours 2–4 and 16–18 over the interval [2019-02-01, 2019-02-28] for 40 EUR/MWh. The simulated payoff distribution for this contract is illustrated on Figure 1.

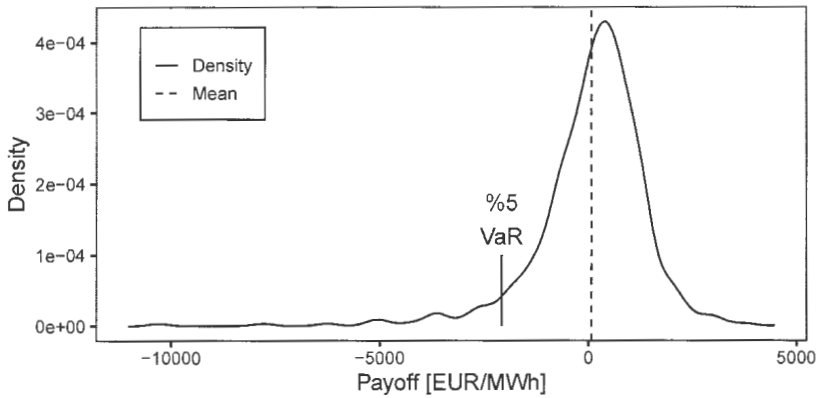


Figure 1: *Simulated payoff distribution for the contract*

The mean of the distribution is 70.98 EUR/MWh, but while the distribution is right-skewed, it exhibits a long left tail. With this simulated distribution, one can calculate risk measures such as Value-at-Risk (VaR), which is simply a quantile in the distribution. The 5% VaR for the example forward contract is -2083.83 EUR/MWh, which means that there is an estimated probability of 0.05 that the contract will constitute a loss at the given size. In Table 3, the 10%, 25%, 75%, and 90% quantiles are listed.

10%	**25%**	**75%**	**90%**
-1370.05	-453.05	826.82	1307.39

Table 3: *Quantiles for the simulated payoff distribution*

4.2 Future considerations

As mentioned above, the first tree of the fitted vine copula consists of primarily Student's t-copulas, so a natural extension would be to see how it compares to a simpler 24-dimensional t-copula and what kinds of tests or criteria can be employed to this end. In addition, since univariate models for base prices are common, one could consider the differences in predicted distributions between the vine-based model and a simpler, univariate model. Finally, it would be interesting to estimate similar models for other major European countries and bidding zones, both for comparison and for collecting into a single, big model.

References

Bedford, T. and Cooke, R. M. (2002). Vines — a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068.

ENTSO-E (2019). ENTSO-E transparency platform. `https://transparency.entsoe.eu`. Last accessed: 2019-08-22.

Escribano, A., Peña, J. I., and Villaplana, P. (2011). Modelling electricity prices: International evidence. *Oxford Bulletin of Economics and Statistics*, 73(5):622–650.

Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *The Canadian Journal of Statistics*, 22(1):47–64.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030 – 1081.

Anders Milhøj

Nyheder i SAS Analytics

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. Der er ikke kommet en opdatering af SAS Analytics for Windows installationer i efteråret 2019. I dette indlæg fokuseres derfor atter en gang på de nye muligheder i denne version 15.1. Sidste års symposieindlæg, Milhøj(2019) koncentrerede sig om bootstrap i PROC TTEST og analyse af kausalstrukturer med PROC CAUSALMED. I år vil muligheder for benyttelse af PROC SGMAP til grafisk visualisering af geodata og en anvendelse af de nye muligheder PROC UCM for analyse af data med overlappende sæsonstrukturer.

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. Denne mulighed for brug af SAS bliver også anvendt af mange af mine bachelorprojektstuderende og specialestuderende.

For yderligere oplysninger om SAS on Demand, se

https://www.sas.com/da_dk/software/on-demand-for-academics.html

I lyset af denne mulighed har jeg aftalt med SAS Danmark, at vi ikke længere omtaler SAS University Edition i min undervisning. Denne mulighed har jeg tidligere omtalt som en mulighed for Mac-brugere i symposieindlæg, men SAS in Demand er enklere i praksis.

SAS og Big Data

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 multi threadet, 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.

SAS Viya

Dette afsnit er direkte coy-pastet fra diverse nyhedsbreve

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. You can write SAS programs using the syntax available for any product that you have licensed and installed.

This week (sidst i november 2019) heralds the release of SAS® Viya® 3.5, which brings new versions of SAS® Visual Statistics, SAS® Optimization, and SAS® Econometrics, along with a brand-new offering of SAS® IML for SAS Viya. Check out our page SAS Analytics: Highlights of SAS Viya 3.5 for details and for links to the What's New documentation. My personal highlights for this new release are PROC SIMSYSTEM for simulating data from Pearson and Johnson distribution families and PROC GAMSELECT for general additive model selection.

Bemærk at SAS Viya altså har overhalet kodeversionerne i SAS med hensyn til opdateringer og at Viay nu også understøtter IML-pakken. På linie med den frie adgang til

SAS via SAS on Demand er der også ved at blive en adgang til SAS Viya for Academics.

Et eksempel på anvendelse af Geodata i SAS

Data i dette eksempel er taget fra en publikation fra Danmarks Meteorologiske Institut i samarbejde med en række rensningsanlæg om regnmåling, DMI(2012). Fra denne publikation er der copy-pastet data fra tabeller over koordinater fra målestationer over hele Danmark, hovedsageligt målestationer nær renseanlæg, samt månedlige nedbørsmængder og oplysninger om dato og nedbørsmængder ved årets mest ekstreme nedbørshændelse.

Oversigtstabellen indeholder oplysninger om en del nedlagte målestationer og der visse fejl i de tabellerne over nedbør, fx på grund af tekniske fejl i måleudstyret, så anvendelserne her baserer sig på et udpluk på lidt over 100 af de ca 150 aktive målestationer.

DMI
Teknisk rapport 13-09

Oversigt over målestationer

Nr.	Nyt Nr.	Navn	Ejer	Zone/ Bælte	N	E
20061	5012	Hjørring Vandværk		32V	6.366.570	560.780
20097	5025	Frederikshavn Materielgård		32V	6.368.560	589.646
20097	5025	Frederikshavn Materielgård	Frederikshavn Forsyning A/S	32V	6.368.560	589.646
20099	5027	Frederikshavn Centralrenseanlæg	Frederikshavn Forsyning A/S	32V	6.366.048	591.707
20211	5047	Sulsted		32V	6.335.760	559.410
20211	5047	Sulsted Stokbrovej Pumpest.	Aalborg Forsyning, Kloak A/S	32V	6.337.114	557.847
20212	5045	Vodskov	Aalborg Forsyning, Kloak A/S	32V	6.329.181	562.128
20298	5049	Gistrup	Aalborg Forsyning, Kloak A/S	32V	6.317.631	560.788
20304	5052	Ålborg Østerport Pumpest.	Aalborg Forsyning, Kloak A/S	32V	6.323.130	557.665
20307	5056	Ålborg Renseanlæg Vest	Aalborg Forsyning, Kloak A/S	32V	6.323.300	552.560
20309	5054	Nørresundby Søvangen Pumpest.	Aalborg Forsyning, Kloak A/S	32V	6.324.730	555.345

Nedbørsmængderne findes i den næste tabel

Nedbør for 2012

Station	Jan	Feb	Mar	Apr	Maj	Jun	Jul	Aug	Sep	Okt	Nov	Dec	År
Nordjylland	73	32	16	58	43	88	89	68	89	78	53	82	778
20097	96	28	17	63	41	94	60	70	87	73	68	77	776
20099	84	26	11	51	41	97	75	71	84	70	65	80	755
20211	76	29	14	60	49	95	101	63	97	72	57	85	799
20212	65	25	14	54	39	89	104	57	87	69	38	47	690
20298	79	33	19	60	47	92	87	65	77	66	41	90	757
20304	64	27	12	47	44	95	111	74	72	63	41	85	736
20307	66	28	18	50	43	75	92	66	78	61	39	90	705
20309	62	24	15	55	52	88	106	84	76	65	41	90	758
20456	64	29	15	46	51	90	95	81	78	80	41	73	742

Tabellen for ekstreme hændelser er

Bilag 2. Oversigt over største nedbørmængde og 10 min. intensitet i 2012 på de enkelte stationer

Station	Navn	Største nedbør-mængde i ét døgn (mm)	Dato	Største nedbør-mængde i én hændel-se (mm)	Dato	Største 10-min intensi-tet µm/s	Dato
20097	Frederikshavn Materialegård	19,8	08/06	19,8	08/06	14,33	06/08
20099	Frederikshavn Centralrenseanlæg	27,8	10/07	26,6	10/07	18,00	06/08
20211	Sulsted Stokbrovej Pumpest.	32,8	06/08	32,6	06/08	17,67	06/08
20212	Vodskov	19,4	06/08	19,4	06/08	7,34	09/07
20298	Gistrup	21,0	06/08	21,0	06/08	11,17	23/08
20304	Ålborg Østerport Pumpest.	37,4	06/08	37,4	06/08	19,00	06/08
20307	Ålborg Renseanlæg Vest	30,6	06/08	30,2	06/08	10,67	06/08
20309	Nørresundby Søvangen Pumpest.	45,8	06/08	45,8	06/08	22,33	06/08

PROC SGMAP

De nye grafiske procedurer i Statistical Graphics suiten (dvs procedurerne med forbogstaverne SG) distribuereres med Base pakken i SAS, så de er til rådighed for alle brugere. Det samme gælder også visse grafiske hjælpeprocedurer fx PROC GPROJECT. Så de kan altså anvendes uden adgang til GRAPH pakken.

PROC SGMAP fungerer som PROC SGPLOT, som plotter (x,y) værdier i et koordinatsystem; forskellen er at (x,y) værdierne er geografiske koordinater i form af

længde- og breddegrader. Desuden kan proceduren indlejre et baggrundsbillede fra de øvrige tilgængelige kort i SAS systemet eller fra andre kilder.

En typisk anvendelse er et kort over regnmålere på Københavns vestegn.

```
ods word file='C:\Users\rkq843\Documents\mini wrk\SAS News
Symp 2020\output.docx';
proc sgmap plotdata=vestegnen;
  openstreetmap;
      scatter x=long y=lat /
    datalabel=station datalabelpos=left
    datalabelattrs=(color=black size=10)
   ;
run;
ods word close;
```

Læg mærke til at outputtet via ods dannes i en Word docx format! Det er en ny facilitet, der endnu er i preproduction. Der er mange faciliteter, der fx danner indholdsfortegnelser etc, se Kelley(2019) for yderligere oplysninger om ODS WORD. På billedet vises resultatet med baggrundskortet "openstreetmap", men der er andre muligheder til fri afbenyttelse, men de egner sig dårligt til symposiebogens trykdesign:

```
esrimap
url='http://server.arcgisonline.com/arcgis/rest/ser-
vices/World_Topo_Map';
```

giver et baggrundskort, der ligner openstreetmap.

```
esrimap
url='http://server.arcgisonline.com/arcgis/rest/ser-
vices/World_Street_Map';
```

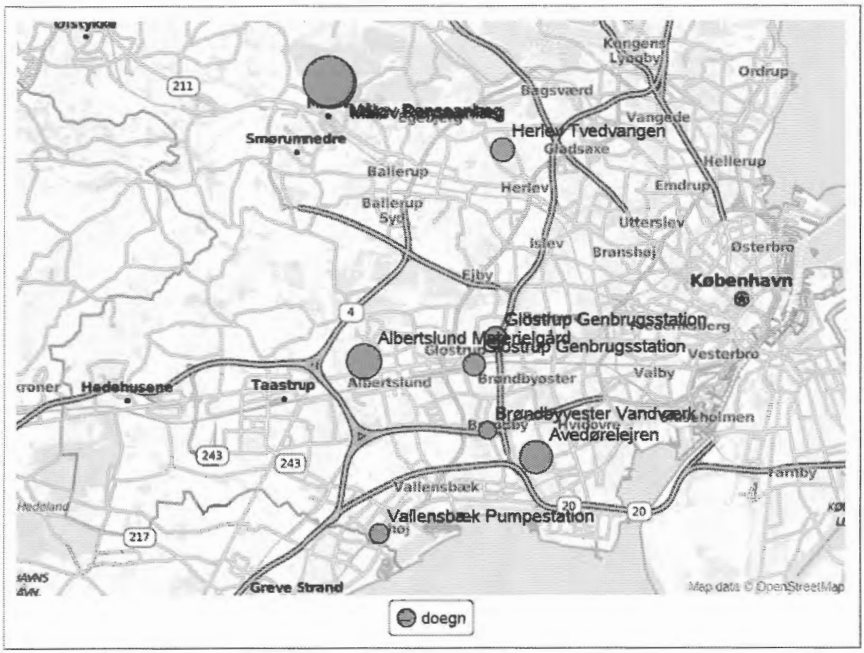
giver et baggrundskort, der også ligner også openstreetmap.

```
esrimap
url='http://server.arcgisonline.com/arcgis/rest/ser-
vices/World_Imagery';
```

giver et sattelitbillede som baggrundsbillede

Nedbørsmængder kan markers på kortet som et boblediagram.

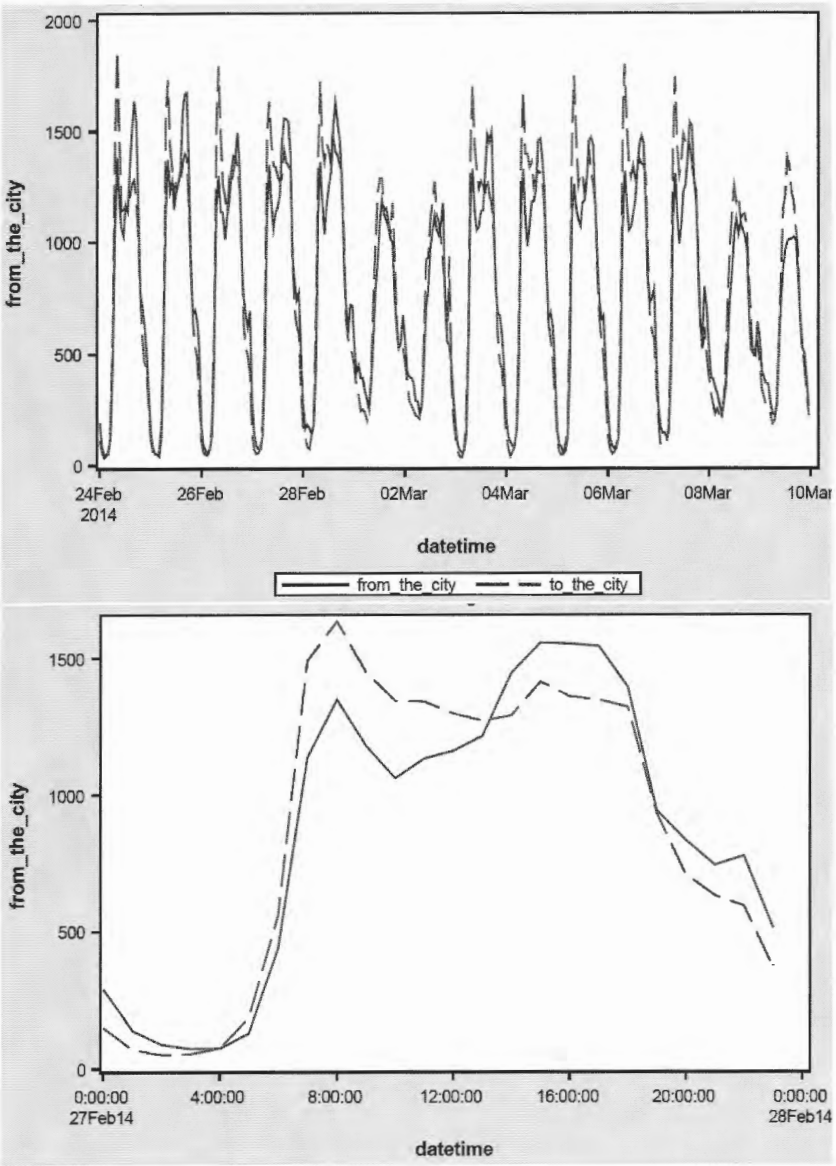
```
proc sgmap plotdata=vestegnen;
  openstreetmap;
  bubble x=long y=lat size=doegn /datalabel=station
     datalabelattrs=(color=black size=10);
run;
```

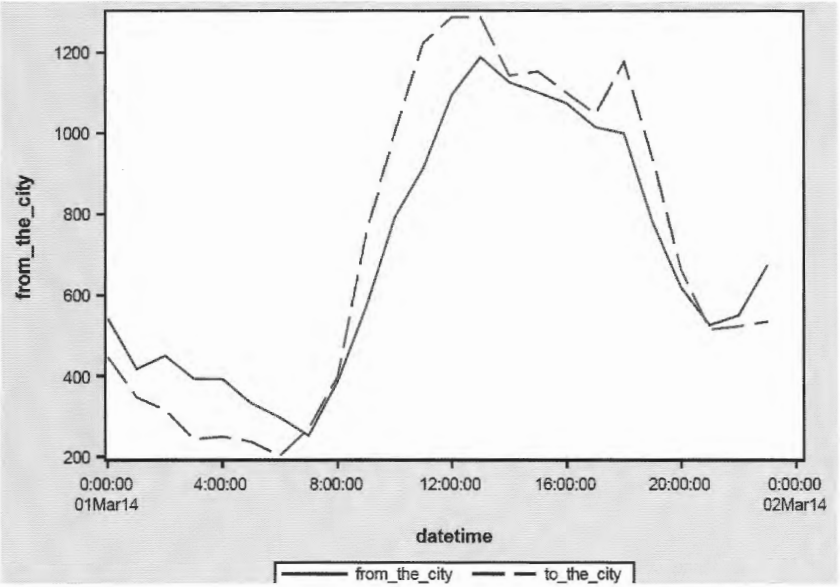


Et eksempel med timedata i PROC UCM

This example uses data on the number of passing bicycles every hour in a period from January 1. at the hour from 00 to 01 (that is the first hour of the new year) and to March 19. Which gives a total of 1872 observations in the winter and early spring 2014. The time series is counted by some automatic device. Such datasets are available for many streets of Copenhagen.

The series is plotted for just two weeks in order to see the details. The period is from Monday February 24. to Sunday March 9. where no special holidays are present. The next two figures are more detailed plots of Thursday February 27. and Saturday March 1'st. The two series on the plots are the number of passing bicycles "to the city" and "from the city". It seems that the number of cyclists to the city is largest in the morning hours while the number of cyclists away from the city are largest in the afternoon hours. This is easily understood as a traffic pattern of people pending to and from work or studies.


2000
1500
1000
500
0
from_the_city
24Feb
2014
26Feb
28Feb
02Mar
04Mar
06Mar
08Mar
10Mar
datetime
from_the_city
to_the_city
1500
1000
500
0
from_the_city
0:00:00
27Feb14
4:00:00
8:00:00
12:00:00
16:00:00
20:00:00
0:00:00
28Feb14
datetime



Saturday and Sunday - however many go for shopping, but a little later in the morning than on ordinary working days. Nightlife especially Friday night is also visible as the numbers of cyclists are large late Friday evening and early Saturday morning.

The basic model is fitted by the code below. The only difference is the seasonal statement where the method is changed from the default dummy parametrization to the parametrization by harmonics and the option print=harmonics gives a table showing the significance of all 168/2 = 84 harmonics. This table is saved as data set named all_harmonics. This dataset is sorted by significance and printed.

```
PROC UCM data=sasts.cycling;
id datetime interval=hour;
model from the city;
irregular p=1;
level plot=smooth var=0 noest ;
season length=168 type=trig print=harmonics var=0 noest;
estimate plot=(panel);
ods output SeasonHarmonics=all harmonics;
run;
proc sort data=all harmonics out=sort;
by descending chisq;
run;
proc print data=sort;
var harmonic period chisq;
run;
```

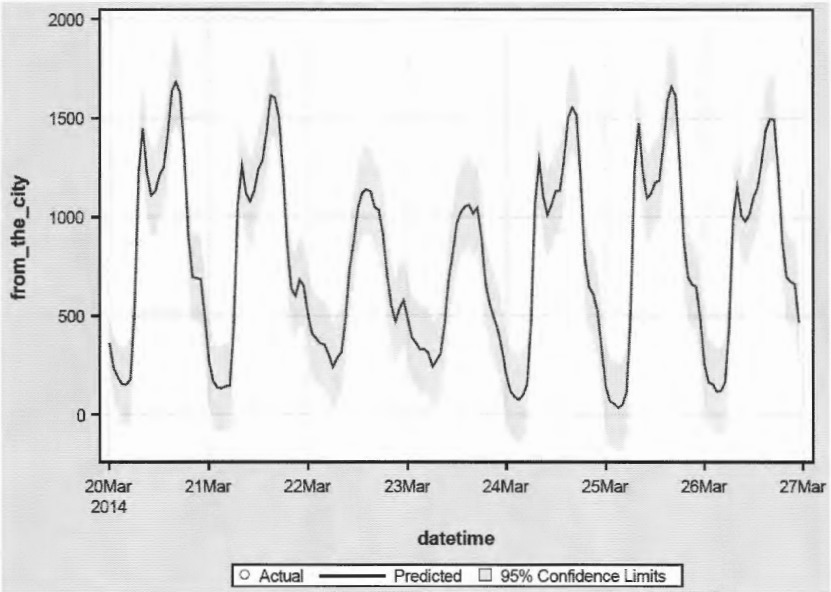
The 25 most important harmonics from the sorted table are shown in the next Table.

Harmonic	period	ChiSq
7	24.00000	8150.51
21	8.00000	3543.76
35	4.80000	2479.17
22	7.63636	494.56
15	11.20000	460.98
49	3.42857	458.12
20	8.40000	439.54
34	4.94118	410.22
13	12.92308	366.59
42	4.00000	301.38
19	8.84211	289.79
16	10.50000	250.63
23	7.30435	240.32
63	2.66667	229.55
14	12.00000	225.08
27	6.22222	207.24
36	4.66667	205.37
6	28.00000	189.73
26	6.46154	169.34
37	4.54054	132.45
5	33.60000	111.51
33	5.09091	110.52
12	14.00000	97.55
28	6.00000	94.58
1	168.00000	92.40

The by far most significant harmonic is 7 which corresponds to period length 168/7 = 24, that is the variation within the 24 hours of a day. The next harmonics are multiples either of 7 or very close to multiples of 7, so they serve to modify the shape of the sinusoid as for instance the amplitude of the daily variation is less in two days of the weekend than for ordinary working days. In the next figure, the number of included harmonics is reduced to these 25 most important harmonics in the table.

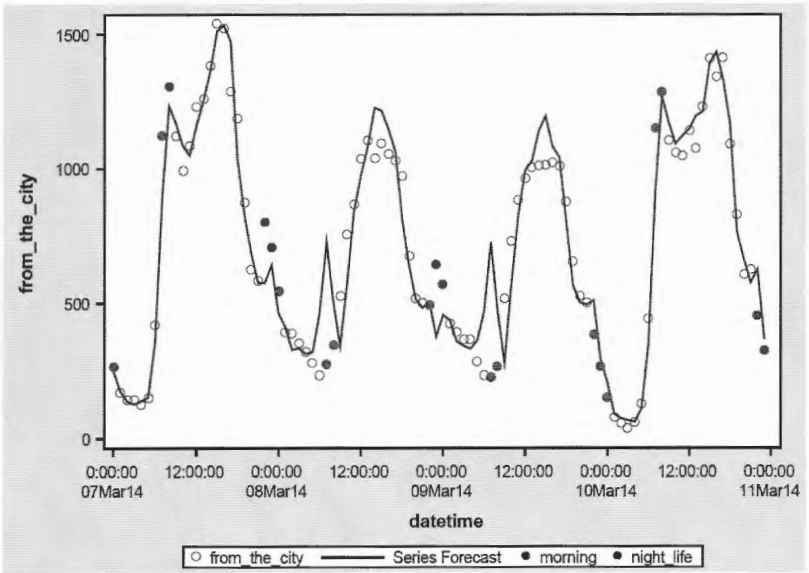
Another way of reducing the number of dummy variables from the 168 potential dummies is to exploit that the 24 hours a day are embedded in the seven days a week. This could be coded rather intuitively by a blockseasonal statement.

The following code includes hourly dummies for the 24 hours a day and daily dummies for seven days a week. These two sets of dummies are then added to the final model. This gives a total of 31 dummies, which is 29 free parameters to be estimated as each set of dummies are restricted to sum zero. All estimated components are saved in the new dataset, named predictions, by the outfor=predictions in the forecast statement.




```
PROC UCM data=sasts.cycling;
id datetime interval=hour;
model from_the_city;
irregular p=1;
level  variance=0 noest;
season length=24 type=trig variance=0 noest plot=smooth;
blockseason nblocks=7 blocksize=24 variance=0 noest
plot=smooth;
estimate plot=(panel) plot=residual;
forecast lead=168 plot=forecasts
outfor=prediction;
run;
```

The reason for the poor fit is clearly seen form the figure above. Here forecasts and observations are plotted for four days, Friday March 7. to Monday March 11., so the effect of a weekend is seen in detail. The red filled dots are the two morning hours 7 and 8 AM where mainly cyclists pass on ordinary work days while very few cyclists pass in these morning hours in the weekend. Moreover the green dots at 10 and 11 PM and also the first hour after midnight make clear that the traffic at night is different the nights after Friday and Saturday compared to the nights before working days



Referencer

Anders Milhøj(2019). Nyheder SAS Analytics 15.1 *Symposium i Anvendt Statistik 2019*, red. Peter Linde, **NAF**

DMI(2013) Drift af Spildevandskomitéens Regnmålersystem – Årsrapport 2012 red. *Teknisk rapport 13-03*, red. Rikke Sjølin Thomsen https://www.dmi.dk/filead-min/Rapporter/TR/tr13-03.pdf

David W. Kelley An Introduction to the ODS Destination for Word
Paper SAS3235-2019
https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceed-ings/2019/3235-2019.pdf

Diverse hjemmesider på support.sas.com.

Primary Care health technology and hospitalizations: the effects of Point-of-care testing of HbA1c on ambulatory care and hospitalizations among type 2 diabetes patients in General practice

Troels Kristensen[1], Kim.Rose Olsen[1] & Christian Volmar Skovsgaard[1]

Corresponding Author: Troels Kristensen
[1] Danish Centre for Health Economics (DaCHE), Department of Public Health, University of Southern Denmark, J.B. Winsløws Vej 9, DK-5000 Odense C, Denmark
Email: Trkristensen@health.sdu.dk

WORK IN PROGRESS: Please do not circulate or cite without permission.

Abstract: Point-of-care testing (POCT) of HbA1c means instant test results and more coherent counseling that may improve diabetes management and reduce ambulatory visits and hospitalizations. From 2008, POCT has been implemented and adopted by a segment of the general practices in the Capital Region of Denmark. The aim of this study is to assess whether the introduction of POCT of HbA1c in general practice (GP) has reduced diabetes-related hospital activity.

We apply difference-in-differences models at the GP clinic level to assess the casual effects of POCT on the following outcomes: 1) hospital ambulatory diabetes care visits, 2) admissions for ambulatory care sensitive diabetes conditions (ACSCs), 3) inpatient admissions for diabetes. The use of POCT is remunerated by a fee, and registration of this fee is used to measure the GP's use of POCT. The control group includes clinics from the same region that did not use POCT. The sensitivity of our results is assessed by an event study approach and a range of robustness tests. We use panel data of 553 GP clinics and approximately 30,000 diabetes patients from the Capital Region of Denmark, observed in the years 2004-2012. Data, based on The Danish Drug Register, The Danish Health Service Register and the National Patient Register, are merged at the patient level using a personal identification number.

There is no significant effect of POCT of HbA1c in GP on hospitalization outcomes. Event study analysis and different treatment definitions confirm the robustness of these results.

This study does not support the hypothesis that POCT in GP reduces diabetes-related hospital activity. However, as hospitalizations are not increased, the use of POCT in GP may be relevant for other reasons, such as improved clinical operations, patient satisfaction, and adherence to medication, or at least warrants further research on other aspects of patient outcomes.

1. Introduction

Timely measurement of HbA1c, which reflects blood sugar levels over the past 3 months, is central to the treatment of T2 diabetes. These measurements are important for both the quality of diabetes care patient outcomes, e.g., the rate of diabetic nephropathy, consumption of health care services and pharmaceuticals, and the subsequent cost of care[1-3]. In general practice (GP) clinics, standard glycemic control is usually performed on venous blood that is sent to a central hospital laboratory where the response is typically available after 1-2 days. Guidelines state that HbA1c should be measured 2-4 times per year depending on how well patients meet glycemic control[4]. However, the guidelines are not met for all T2 diabetes patients[5]. This lack of adherence may arise for reasons such as poor self-management and subsequent missing alignment to guidelines[6, 7]. Poor self-management may be related to such issues as reduced ability to receive care due to comorbidities and multimorbidity (e.g., psychiatric diseases) and social problems[8].

One approach to increase compliance and improve outcomes may be to introduce point-of-care tests (POCT) of HbA1c in GP clinics[9]. This is an alternative to standard hospital laboratory tests where the GP performs finger-stick capillary blood collection[10]. The patient receives the test results immediately during the consultation (within 5-7 min). Thus, the GP clinic shortens the time to treatment decision-making and improves diabetes management – e.g., via increased compliance[11]. However, this technology also changes the behavior of patients and healthcare providers and allows for optimization of processes that can lead to better health, lower cost and economic benefits from a societal perspective[12]. In this paper, we focus on patient outcomes. Switching from laboratory testing to POCT may improve medical outcomes for patients. One way is that POCT may increase compliance with HbA1c testing among patients with an elevated risk of non-attendance. For instance, vulnerable and complex patients with social problems that the GP would normally not send to an HbA1c test at a laboratory or ask for a revisit[13, 14]. Hence, POCT reduces the demand for self-discipling (e.g., in terms of fewer required revisits)[15]. Another contribution is that POCT may lead to timely review of results and therapy without delays because the care can be delivered immediately without revisits after an external laboratory test[16]. A third possible effect includes a higher level of patient satisfaction and greater patient loyalty as a result of the practice's reputation for applying state-of-the-art technology[12]. Furthermore, it has been argued that POCT catalyzes greater engagement with the patient and better adherence to medication and compliance to diabetes management in general[17]. Finally, performing the POCT analysis close to the patient reduces the demand for hospital laboratory capacity and transportation of blood samples. Thus, the potential of POCT technology

goes beyond individual patient health and affects clinical decision-making and service planning, which may also have positive impacts on patient outcomes. The short-term intermediate outcome is expected to be improved HbA1C levels and subsequently fewer hospital encounters due to improved control via POCT of HbA1c[18]. Specifically, the association between glycemic levels and hospitalizations is well established[19, 20]. Therefore, we hypothesize that patients with diabetes monitored in GP clinics via POCT on average have fewer diabetes-related hospital ambulatory visits and hospitalizations. This is in line with calls for more integrated and value-based (outcome) approaches to evaluate the diagnostic technologies[21]. Accordingly, we follow [22, 23] and use hospitalizations and diabetes-related ambulatory visits to measure diabetes outcomes. The aim of this study is to assess the effect of POCT of HbA1c on hospital ambulatory visits and hospitalizations from 2008-2012 through a natural experiment in Denmark.

2. Methods and data

Ultimo 2008, it was agreed that general practitioners in the Capital Region should be paid a new fee (of Danish Krone 115,97) per POCT of HbA1c. The uptake of POCT by some GPs and lack of implementation by other GPs in the region represents a natural experiment[24]. We applied a difference-in-differences (DID) method to analyse the natural experiment and estimate the effects of POCT on average rates of ambulatory visits and hospitalizations[25]. POCT clinics were defined as clinics using POCT at least once during a year. Both differences in ambulatory visits and hospitalizations before and after were calculated between patients listed in POCT clinics compared to GP clinics that only used a central hospital laboratory or walk-in labs to measure HbA1c. The special fee-for-service fee that was introduced for POCT of HbA1c ultimo 2008 was applied as proxy measure for POCT of HbA1c among diabetes patients in general practice[26].

2.1 Treatment, control and introduction years: The treatment group was defined as clinics that introduced POCT during the period 2009-2012 and used it through 2012. Table 1 below shows the treatment indicator variable I_{it} for all combinations of possible introduction years (T_i) and years of observation (t_i). This flexible approach allowed us to include GP clinics as they introduced the technology. The GPs in the Capital Region that used a central hospital laboratory test or walk-in labs and never used POCT were applied as the control group. Hence, the treatment and control group are exposed to the same "competing" regional explanations for change in outcomes.

Table 1: Indicator of treatment I_{it} for T_i (introduction year) and year t_i

		Year (t_i)						
	I_{it}	2006	2007	2008	2009	2010	2011	2012
Int year (T_i)	2009	0	0	0	1	1	1	
	2010	0	0	0	0	1	1	1
	2011		0	0	0	0	1	1
	2012			0	0	0	0	1

Note: The table shows for each year t_i whether a clinic that introduced POCT in year T_i is included in the treatment group.

The treatment measure (C_{it}) was defined as the share of diabetes patients linked to the GP clinic who were treated with POCT in a given year. This approach makes it feasible to account for the intensity in the use of POCT, which is expected to be positively related to the treatment effect.

2.2 Identification and estimation strategy: To identify the average treatment effect (ATT) of POCT on ambulatory visits and hospitalizations and to account for selection bias, we use a DID framework with a continuous treatment variable and include GP and year fixed effects as well as time-varying control variables. Hence, we estimate the effect of POCT on outcomes for GP clinics that in a given year used POCT with clinics that did not. To take into account GP variation unrelated to the introduction of POCT, we included time fixed effects. Thus, we estimate the following fixed effect regression model[27-29]:

$$y_{it} = \delta C_{it} I_{it} + \beta x_{it} + \theta_t + GP_i + \mu_{it} \qquad (1)$$

where y_{it} represents our outcome measures of hospital activity for clinic $i = 1, \dots, N$ in year $t = 2006.., 2012$, C_{it} is a continuous treatment variable measuring the share of diabetes patients at GP_i treated with POCT in year t, I_{it} is a shift-variable which turns to one in the flexible introduction year (2009-12) in which GP_i introduces POCT, and hence, $C_{it} I_{it}$ is the interaction of interest representing the effect of treatment of POCT after the introduction. x_{it} is a vector of observed time-varying control variables, θ_t represents time/year fixed effects that captures time trends common to both treatment and control groups. GP_i captures GP clinic level time-invariant unobserved effects (i.e., GP fixed effects).

2.3 Event study and sensitivity analysis: To visualize our findings and test the sensitivity, we used an event study approach that investigates and visualizes the yearly effect of POCT in the years before (falsification) and after (treatment) the flexible introduction

year[27]. For each GP, we define the variable k_i, which measures the individual number of years after the flexible introduction year of POCT. For years before the introduction, k_i is negative (see Table 2 below). When $k < -1$, the parallel assumption is tested in the sense that after taking the covariates into account, we should see no difference in outcomes (relative to the baseline year) between the treatment and the control group prior to the introduction.

Table 2 Years before and after the flexible introduction year of POCT

		Year (t_i)						
	k_i	2006	2007	2008	2009	2010	2011	2012
Int year (T_i)	2009	-3	-2	-1	0	1	2	
	2010	-4	-3	-2	-1	0	1	2
	2011		-4	-3	-2	-1	0	1
	2012			-4	-3	-2	-1	0

Note: The table shows for each year t_i the number of years k_i before and after the introduction of POCT in the clinic. When t_i is the same as the introduction year, then $k_i = 0$.

The event study and sensitivity analyses were conducted through regression equation (2) which estimates the effects for the years before and after the introduction of POCT:

$$y_{it} = \alpha + \sum_{k=k_i^{min}}^{-2} \beta_k^{Before} C_{it} \mathbf{1}[t_i - T_i = k_i] + \sum_{k=0}^{k_i^{max}} \beta_k^{After} C_{it} \mathbf{1}[t_i - T_i = k_i] + \gamma x_{it} + \theta_t + GP_i + \mu_{it} \quad (2)$$

The variables $y_{it}, C_{it}, x_{it}, \theta_i, GP_i$ were defined for equation (1), $T_i \in [2009; 2012]$ denotes the year where each GP_i introduces POCT, and $t_i - T_i = k_i \in$ [-4;2] in Table 2 represents the number of years before and after GP_i introduced POCT. Therefore, the definition of what is before and after becomes dynamic. The range for $k_i^{min} \in$ [-4;-3] and $k_i^{max} \in$ [0;2] depend on the year of introduction. The indicator function $\mathbf{1}[\cdot]$ is one if $t - T_i$ equals k (shown in Table 2) and zero otherwise. Thus, β_k^{Before} in the first summation in (2) estimates the difference in outcomes between control and treatment for each year (k) *Before* the introduction of POCT, and β_k^{After} in the second summation in (2) estimates the effect of POCT on outcomes for each year (k) *After* the introduction of POCT at the GP level.

The reference year is always the year before GP_i introduces POCT, $k = -1$. This year (and not $k = 0$) was used to ensure that the GP did not use the equipment in the reference

year. All estimates should be interpreted relative to that flexible reference year. As GP_i introduces POCT at some point throughout the year $k = 0$, β_0^A should be interpreted in that context; effects might not have occurred fully yet.

2.4 Robustness tests: To investigate the robustness of our results, we used a) an alternative dichotomous treatment measure instead of the continuous measure, b) used a treatment definition that only included the first mover clinics that introduced POCT in 2009, and c) performed subgroup analysis for different sets of patient types.

The alternative dichotomous treatment measure defined GP clinics as treated if they used at least 5 POCTs during a year in all subsequent years after the clinic introduction rather than the continuous measure. The model of the early adopters included the clinics that introduced POCT in 2009 and used it throughout the entire period of observation. This test was performed to explore the effects of POCT among patients of the early adopters, as these clinics used POCT more intensively and for the longest period. These GP clinics were also likely to be those that have a special interest in diabetes and hence might constitute a potential selection problem in our main model. Finally, subgroup analyses for selected types of patients (high/low Charlson Index, high/low educational level, above/below age 65, and Danish/foreign ethnicity) inspired by a recent study were undertaken to explore the robustness of the results across these subgroups[26].

2.5 Ambulatory and hospitalization measures: The treatment effect of POCT on diabetes patients was assessed for three GP clinic level outcome variables: 1) average hospital outpatient diabetes rates, 2) average inpatient diabetes hospitalization rates and 3) average ACSC diabetes hospitalization rates. ACSC admissions were added because ACSC admissions are widely used as indicators of primary care outcome[30].

2.6 Covariates: This study controls for potential selection bias via time-varying covariates to obtain a conditional common pretreatment trend in ambulatory visits and hospitalizations since the implementation of POCT was voluntary. These covariates are also used to adjust for differences between the control and the treatment group of GP clinics around the time POCT was implemented. The vector of control variables includes three subsets of clinic covariates: a) the diabetes management characteristics of GP clinics (the clinics' proportion of diabetes patients, the share of patients registered on a special diabetes bundle payment fee, the average number of GP visits and the proportion of the diabetes patients with a diabetes experience (diabetes age > 5 years)); b) the average GP clinic level morbidity burden of all patients attached to the clinic (the Charlson index); and c) the socioeconomic characteristics of the GP clinic patients (the proportion of elderly, unemployed, singles, and average family income). These socioeconomic characteristics are assumed to capture the developments in the morbidity

burden that are not captured by the Charlson index. The clinic-level diabetes management characteristics a) were included as proxies for GPs with special interest in and knowledge about diabetes. The average Charlson index (excluding diabetes) and the proportions of socioeconomic characteristics were used to adjust for dynamic differences in patient morbidity burden and socioeconomic background. Diabetes was excluded from the Charlson index to avoid endogeneity regarding the outcomes of interest. To explore the contribution from each set of covariates and combinations of these three subsets, a-c) were included in three different ways: Model 1 includes the key term of interest (the treatment interaction term), Model 2 adds covariates from (a) and Model 3 adds covariates from (b) and (c) to adjust for the above-mentioned potential selection bias.

2.7 Defining the included T2 diabetes patients: We use a panel data set covering the years 2004-2012 for a cohort of T2 diabetes patients from the Capital Region of Denmark. The cohort of T2 diabetes patients was defined by the algorithm of the Danish Diabetes Register based on The Danish Drug Register, The Danish Health Service Register and the National Patient Register. Patients were required to be above 18 years of age, alive in 2012, and living in Denmark. Furthermore, at least one out of the following three criteria had to be met in a given year: 1) The patients had redeemed at least one prescription for anti-diabetic drugs with ATC code A10A* or A10B* (*: Including subgroups). ATC-code A10BA02* (metformin) was excluded for women between the ages of 20-40 (gestational diabetes). 2) The patients had received at least three blood sugar or HBA1c tests from their GP or a specialist in the primary sector 3) The patients were registered with one of the following ICD10 codes in the hospital sector: DE11, DE12, DE13, DE14, DO24, or DH360. Patients who in any given year received medicine related to T1 only were excluded (see appendix 1 and 2 for further details).

To be included in the cohort, patients had to be identified by the algorithm described above at least once during the period 2004-2006. The cohort was identified in the period leading up to the analysis period 2006-2012 to allow us to understand and explore the difference in the control and treatment group before and after the intervention without the preintroduction period being influenced by accession and attrition. Similar algorithms used to identify patients with type 2 diabetes have been shown to include patients without diabetes. To address this issue, patients who were identified only once during the entire period 2004-2012 were excluded. In total, the cohort consisted of more than 30,000 T2 diabetes patients. Each patient was linked to the GP clinic they used most frequently. Data approval for conducting the study was provided by the Data Protection Agency (ref. nr 17/6021). An anonymized id number using public health

services was used to merge data at the individual level from the Danish administrative registers.

3. Results

3.1 Descriptive clinical characteristics: Table 3 shows the mean values for the outcome variables and time-varying covariates for the treatment group (231 clinics) and the control group (322 clinics) at the clinic level in 2008 before the introduction of POCT.

Table 3 Descriptive clinic characteristics

	2008		
	Mean		P-value
	Control (1)	Treatment (2)	Difference (3)
Time-varying controls			
Diabetes characteristics & management:			
Proportion of diabetes patients	0.023	0.028	**0.000**
Share of clinics, bundle payments	0.044	0.062	0.422
Average # GP consultations	6.770	6.998	0.212
Share of diabetes patients age > 5	0.544	0.515	**0.011**
Morbidity burden:			
Charlson index, (excl. diabetes)	0.208	0.199	0.415
Socioeconomic proportions:			
Proportion of elderly patients, age> 65	0.502	0.490	0.330
Proportion of unemployed patients	0.042	0.039	0.495
Patient family income	204,13	204,816	0.894
Proportion of single patients	0.394	0.371	**0.044**
N	322	231	

Notes: Descriptive statistics for the treatment and control groups before (2008), the introduction of the framework for POCT. The treatment group includes clinics that became POCT clinics during the period from 2009 to 2012. Group differences significant at the 5% level are in bold.

In the treatment group (POCT clinics), the average proportion of diabetes patients was 2.8%, the share of clinics using bundle payment was 6.2%, the yearly average number of GP consultations was 7, and the share of patients with a diabetes age above 5 years was 52%. The clinic level morbidity burden (Charlson index excl. diabetes) was on average 0.199, and the socioeconomic proportions were as follows: 49.0% elderly above 65 years, 3.9% unemployed and 37,1% living alone (single). The patients had an average

family disposable income of DKK 204.816. Difference between treatment and control groups: The control group had a lower proportion of diabetes patients (2.3%), a higher share of patients with diabetes age above 5 (54.4%), and a higher share of single patients (39.4%).

3.2 The effect of POCT of HbA1c: The results for the models in Table 4 show the effect of POCT of HbA1c on diabetes-related admissions (columns 1-3), on diabetes-related ACSC admissions (coloumns 4-6) and on ambulatory care activity (coloumns 7-9).

Table 4 Estimation of POCT treatment effects via DID models & fixed effect regression

Dependent variable:	Diabetes Admissions			ACSC Diabetes Admissions			Ambulatory Care, Diabetes		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
POCT $*1^{year}$	0.00957	0.00863	0.00999	-0.00198	-0.00220	-0.0023	-0.0152	-0.0158	-0.0131
	(0.70)	(0.62)	(0.75)	(-0.62)	(-0.68)	(-0.71)	(-0.87)	(-0.90)	(-0.75)
Diabetes:									
Proportion of diabetes patients		1.116	1.075		0.315	0.317		1.185	1.439
		(1.30)	(1.23)		(1.64)	(1.64)		(0.88)	(1.11)
Share of bundle payments		0.000622	0.00232		0.00205	0.0022		0.0109	0.0123
		(0.09)	(0.35)		(0.99)	(1.06)		(1.25)	(1.41)
# GP consultations		0.00335	0.00256		-0.000140	-0.00019		-0.0021	-0.00226
		(1.52)	(1.23)		(-0.28)	(-0.36)		(-0.67)	(-0.79)
Share of diabetes patients age>5		-0.0366	-0.055		-0.00507	-0.00625		-0.0866	-0.0864
		(-0.45)	(-0.71)		(-0.30)	(-0.37)		(-0.60)	(-0.59)
Comorbidity:									
Charlson, index (excl. diabetes)			0.188***			0,0101			0.215***
			(4.10)			(1.49)			(6.30)
Socioeconomic:									
Prop. of elderly patients, age>65			0.05			0.00797			-0.0498
			(1.04)			(0.82)			(-0.60)
Proportion of un-employed patients			-0.00159			0.0121			-0.114
			(-0.02)			(-0.73)			(-1.09)
Patient family income			1.29e-			-1,8E-08			1,6E-08
			(-3.10)			(-1.65)			(0.19)
Proportion of single patients			0.0727			-0,00601			-0.08
			(1.47)			(-0.48)			(-1.10)
Year & GP FE	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes
Mean outcomes,08	0.100	0.100	0.100	0.011	0.011	0.011	0.348	0.348	0.348
R^2	0.293	0.294	0.338	0.077	0.077	0.080	0.643	0.643	0.657
N	3871	3871	3871	3871	3871	3871	3871	3871	3871

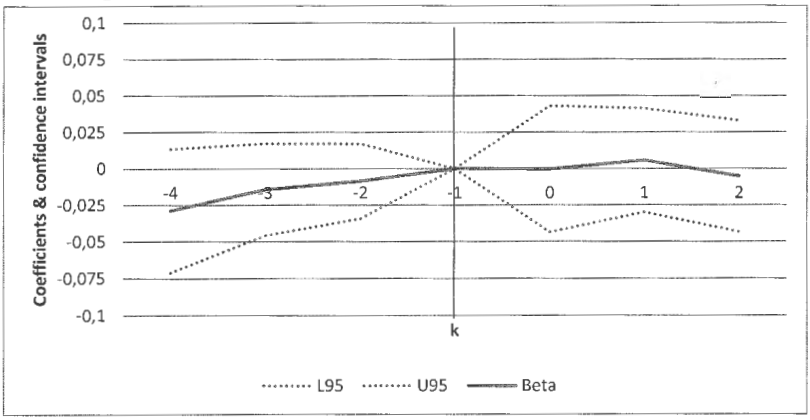
*Notes: The table shows the effect of POCT of HbA1c on diabetes-related admissions, ACSC admissions and ambulatory care visits. POCT $*P^{year}$ is the interaction between the treatment variable POCT and $P^{year,}$ which is one from year of introduction onwards and zero before. Year and clinic fixed effects are included in all models. Control variables are all at the level of the clinic. T-statistics in parentheses. Significance levels: * $p<0.05$, ** $p<0.01$ and *** $p<0.001$*

The results in Table 4 do not support the hypothesis that POCT should reduce hospitalizations and/or ambulatory care visits in the secondary sector. For all models, none of the estimated treatment effects are significant. Covariates that vary between the treatment and control group are likely to be significant. The majority of the time-varying

covariates show limited impact on the outcomes in the DID model. The only covariate that explains the differences in outcomes between the treatment and control groups after the introduction of POCT is the Charlson index. However, if we compare the full model with the restricted model, the inclusion of the Charlson index does not alter the significance or the magnitude of the effect of POCT.

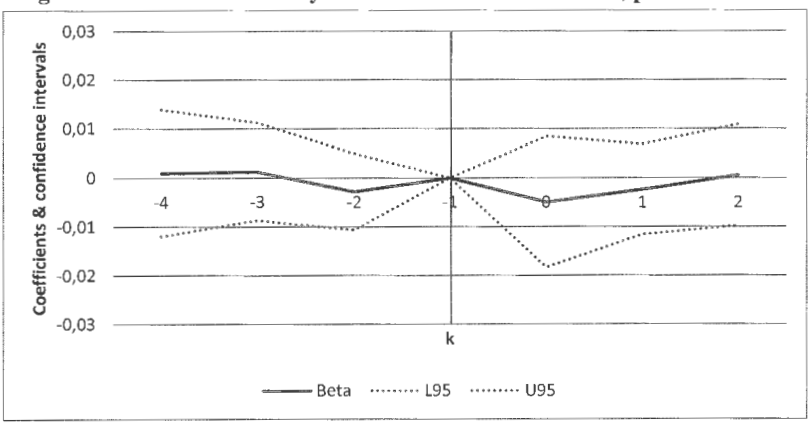
3.3 Event study analysis: The results in Figure 3, Figure 4 and Figure 5 show the estimated point estimates of the effect of POCT on the ambulatory care visits, standard admissions and ACSC admissions, respectively.

Figure 3 Relative event study effects on admissions, point estimates.



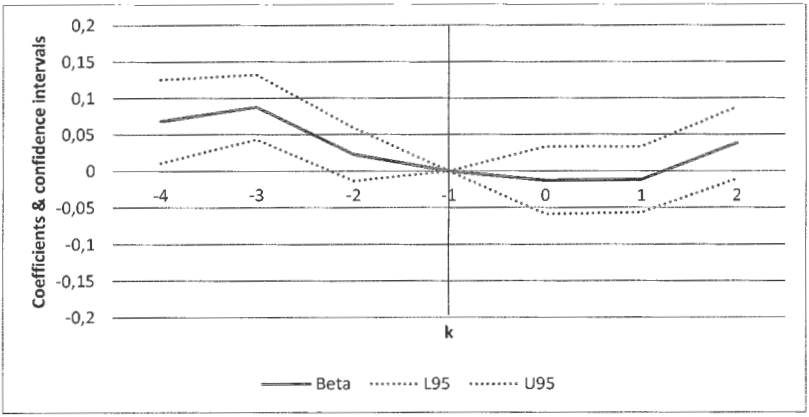
Notes: Year and clinic fixed effects are included in all models, and the full set of covariates from Table 4 are included in all estimations. The dotted lines represent 95% confidence intervals.

Figure 4 Relative event study effects on ACSC admissions, point estimates.



Notes: See Figure 3

Figure 5 Relative event study effects on ambulatory care, point estimates



Notes: See Figure 3.

Parallel assumption: For both types of hospitalizations, the parallel assumption seems to be fulfilled. The justification is that we see no difference in outcomes (relative to the baseline year) between the treatment and the control group prior to the POCT introduction after taking into account covariates. For ambulatory care visits, there are minor exceptions for single years ($k = -4$ *and* $k = -3$). This means that 4 and 3 years prior to the introduction, patients from the POCT clinics had more ambulatory care visits in contrast to the remaining preintroduction years.

Treatment effect: For $k > 0$, a positive estimate indicates a treatment effect in line with the hypothesis. The graphs in Figure 3, Figure 4 and Figure 5 show the effect over time, so we can investigate whether, for instance, an early effect declines over time or vice versa. In most cases, we see no sign of any particular patterns over time. For ambulatory care, we see a slight increase in the second year after introduction. However, as shown in Table 4, where we look at the average post- and preintroduction periods in the model, the effect on ambulatory care is negative and insignificant.

3.4 Robustness: Several robustness tests were performed. First, we conducted subgroup analysis on patients with above/below median Charlson Index, educational level, age, and on Danish/foreign background. These subgroup sensitivity analyses did not show treatment effects of POCT of HbA1c on hospital activity for any of the eight subgroups (results available upon request). Second, we tested a binary treatment definition. The results were similar when using a binary treatment outcome instead of a continuous one. This indicates that the results are not driven by the choice of treatment measure. Finally, we tested a model that only included GP clinics in the treatment group that introduced

POCT in the first possible year (2009). This model revealed a borderline significant reduction in ambulatory care due to the use of POCT when not all covariates were included. However, the parallel assumption was not fulfilled for this specific group of clinics, which seemed to be driving that result.

4. Discussion

The results of this natural experiment do not support the hypothesis that POCT of HbA1c leads to a reduced number of visits and hospital admissions. At the same time, the results also indicate that the use of POCT of HbA1c in GP clinics does not harm patients in terms of admission rates. This finding is important since the technology has already been argued to reduce operations cost and improve patient satisfaction and has been recommended by recognized stakeholders such as the American Diabetes Association. The finding of no impact on hospital activity is robust to a number of robustness checks and sensitivity analyses. The results only reveal indications of a reduction in outpatient activity for the subgroup analysis of the early adopter clinics that had been using POCT for all four subsequent years. This could indicate that an effect takes longer time to materialize than our follow-up period allows and/or that the early adopters are different.

Selection bias: By using a fixed population of patients with diabetes defined in the preanalysis period 2004-2006 before the natural experiment, we eliminate the problem of patient selection into the treatment group in the later phases of the experiment. However, GPs that choose to use POCT may be subject to self-selection. An obvious reason could be that GPs in POCT clinics are more interested in diabetes and have better performing diabetic patients due to the GPs being more competent with diabetes management. The latter assumption is reasonable if we anticipate that the use of new technology such as POCT is a good proxy for treatment quality. This means that a negative treatment effect may be due to selection rather than the effectiveness of POCT. However, the event study in this paper does not indicate that selection is a problem after controlling for a number of time-varying observable covariates as well as time-invariant unobservable variables controlled for by the fixed effects. In contrast, it seems that outpatient admissions are higher in the POCT clinics before treatment, indicating a reverse selection compared to the expectation of lower hospital utilization in the POCT clinics. This selection bias may work in the reverse direction and potentially eliminate the hypothesized effect of POCT of HbA1c on outpatient admissions. Hence, if we assume that we have adjusted for severity of the disease, the negative sign of the treatment effect for the ambulatory admissions may, if anything, be underestimated due

to remaining selection in the model. This interpretation seems to be supported by the indications of negative significant treatment effects for the subgroup analysis of the GPs who have been using POCT for 4 years.

The absence of the implementation of the POCT of HbA1c in the four other regions has been analyzed and discussed by health care professionals (Kristensen et al., 2017). However, the regulator's arguments for lack of implementation have not been explicitly communicated. Regional decisions regarding introduction may, for instance, be related to: a) medical opinions in the individual regional laboratory committees or other regional preferences regarding the use of hospital laboratories capacity, b) the transportation capacity between clinics and laboratories and c) the preferences for central testing and analysis, for instance, to ensure the quality of tests and that the test results are registered centrally for R&D purposes. We expect that this study can help the regions to an improved understanding of the effect of POCT of HbA1c technology and motivate more explicit communication regarding these decisions.

Limitations: This study hypothesized fewer diabetes-related hospital ambulatory visits and hospital admissions for clinics using POCT. Nevertheless, some professionals in the field doubt that POCT results in better treatment outcomes. This is due to ambiguous evidence and concerns regarding the effects, the right implementation and appropriate use of the technology[31]. For instance, insufficient training of staff in using the equipment and lack of appropriate calibration may contribute to inappropriate use. We cannot rule out that challenges related to the appropriate implementation of POCT may be part of the explanation why this study did not confirm the hypothesis.

The time horizon of the analysis from 2006 to 2012 is relatively short. Diabetes is a chronic and slowly progressive disease. It may be optimistic to expect an effect on hospital admission rates within a time frame of up to four years. However, other studies on diabetes management tend to find reductions in hospitalization rates within the same timeframe [22]. In this light, we find the present timeframe to be useful for assessing the role of POCT on hospital admissions.

To understand the role of POCT at the patient-health system interface, more research is needed to ascertain the effects on prescription behavior, outcomes, distributional implications across socioeconomic groups of patients and how the policy environment and health care practices influence the usefulness of POCT[13]. For instance, the policy environment has changed recently. In 2017, even more treatment of selected chronic conditions (COPD and T2 diabetes) were shifted to general practice, and in 2018, a new fixed comprehensive incentive structure was implemented for type 2 diabetes patients in Danish general practice.

5. Conclusion

This study does not support the hypothesis that POCT of HbA1c in general practice reduces diabetes-related hospital activity. The results indicate that POCT of HbA1c has no effect on hospital activity. Hence, if the implementation of POCT of HbA1c improves patient satisfaction and clinical operations via fewer visits and improved management in one operation, it may be worthwhile to implement POCT of HbA1c in general practice or at least warrant further research on the impact of HbA1c on other aspects of patient outcomes.

Literature

1. Bonke, F.C., Donnachie, E., Schneider, A., & Mehring, M. (2016). Association of the average rate of change In HbA1c with severe adverse events: a longitudinal evaluation of audit data from the Bavarian Disease Management Program for patients with type 2 diabetes mellitus. *Diabetologia,* 59, 286-293.
2. Kostev, K., Rockel, T., & Jacob, L. (2016). Impact of Disease Management Programs on HbA1c Values in Type 2 Diabetes Patients in Germany. *J Diabetes Sci Technol.*
3. Little, R.R., & Rohlfing, C.L. (2013). The long and winding road to optimal HbA1c measurement. *Clin ChimActa,* 418, 63-71.
4. American Diabetes, A. (2017). Standards of Medical Care in Diabetes-2017 Abridged for Primary Care Providers. *Clin Diabetes,* 35, 5-26.
5. Styregrupperne, & kompetencecenterne. (2013). Dansk Diabetes Database: National årsrapport 2012, 1.marts 2012 – 28. februar 2013. Kommenteret version 1.0. Regionernes Kliniske Kvalitetsudviklings-program.
6. Ellis, J., Boger, E., Latter, S., Kennedy, A., Jones, F., Foster, C., et al. (2017). Conceptualisation of the 'good' self-manager: A qualitative investigation of stakeholder views on the self-management of long-term health conditions. *Soc Sci Med,* 176, 25-33.
7. Savoca, M.R., Miller, C.K., & Quandt, S.A. (2004). Profiles of people with type 2 diabetes mellitus: the extremes of glycemic control. *Soc Sci Med,* 58, 2655-2666.
8. Larsson, A., Greig-Pylypczuk, R., & Huisman, A. (2015). The state of point-of-care testing: a European perspective. *Ups J Med Sci,* 120, 1-10.
9. Price, C.P., & St John, A. (2019). The value proposition for point-of-care testing in healthcare: HbA1c for monitoring in diabetes management as an exemplar. *Scand J Clin Lab Invest,* 79, 298-304.
10. Schols, A.M.R., Dinant, G.J., Hopstaken, R., Price, C.P., Kusters, R., & Cals, J.W.L. (2018). International definition of a point-of-care test in family practice: a modified e-Delphi procedure. *Fam Pract,* 35, 475-480.
11. Ang, S.H., Thevarajah, M., Alias, Y., & Khor, S.M. (2015). Current aspects in hemoglobin A1c detection: a review. *Clin Chim Acta,* 439, 202-211.
12. Patzer, K.H., Ardjomand, P., Gohring, K., Klempt, G., Patzelt, A., Redzich, M., et al. (2018). Implementation of HbA1c Point of Care Testing in 3 German Medical Practices: Impact on Workflow and Physician, Staff, and Patient Satisfaction. *J Diabetes Sci Technol,* 12, 687-694.
13. Haenssgen, M.J., Charoenboon, N., Althaus, T., Greer, R.C., Intralawan, D., & Lubell, Y. (2018). The social role of C-reactive protein point-of-care testing to guide antibiotic prescription in Northern Thailand. *Soc Sci Med,* 202, 1-12.
14. Rogvi, S., Tapager, I., Almdal, T.P., Schiotz, M.L., & Willaing, I. (2012). Patient factors and glycaemic control-associations and explanatory power. *Diabet Med,* 29, e382-389.
15. Petrakaki, D., Hilberg, E., & Waring, J. (2018). Between empowerment and self-discipline: Governing patients' conduct through technological self-care. Soc Sci Med, 213, 146-153.
16. St John, A., & Price, C.P. (2013). Economic Evidence and Point-of-Care Testing. *Clin Biochem Rev,* 34, 61-74.

17. Gialamas, A., Yelland, L.N., Ryan, P., Willson, K., Laurence, C.O., Bubner, T.K., et al. (2009). Does point-of-care testing lead to the same or better adherence to medication? A randomised controlled trial: the PoCT in General Practice Trial. *Med J Aust,* 191, 487-491.
18. Wolters, R.J., Braspenning, J.C.C., & Wensing, M. (2017). Impact of primary care on hospital admission rates for diabetes patients: A systematic review. *Diabetes Res Clin Pract,* 129, 182-196.
19. Gall, M.A., Borch-Johnsen, K., Hougaard, P., Nielsen, F.S., & Parving, H.H. (1996). [Albuminuria and glycemic control. The significance for mortality in non-insulin-dependent diabetes mellitus]. *Ugeskr Laeger,* 158, 6907-6911.
20. Menzin, J., Korn, J.R., Cohen, J., Lobo, F., Zhang, B., Friedman, M., et al. (2010). Relationship between glycemic control and diabetes-related hospital costs in patients with type 1 or type 2 diabetes mellitus. *J Manag Care Pharm,* 16, 264-275.
21. Price, C.P., Wolstenholme, J., McGinley, P., & St John, A. (2018b). Translational health economics: The key to accountable adoption of in vitro diagnostic technologies. *Health Serv Manage Res,* 31, 43-50.
22. Dusheiko, M., Doran, T., Gravelle, H., Fullwood, C., & Roland, M. (2011). Does higher quality of diabetes management in family practice reduce unplanned hospital admissions? *Health Serv Res,* 46, 27-46.
23. Iezzi, E., Lippi Bruni, M., & Ugolini, C. (2014). The role of GP's compensation schemes in diabetes care: evidence from panel data. *J Health Econ,* 34, 104-120.
24. Craig P, Gibson M, Campbell M, Popham F, Katikireddi SV. Making the most of natural experiments: What can studies of the withdrawal of public health interventions offer Prev Med. 2018;108:17-22
25. Craig, P., Cooper, C., Gunnell, D., Haw, S., Lawson, K., Macintyre, S., et al. (2012). Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. *J Epidemiol Community Health,* 66, 1182-1186.
26. Kristensen, T., Waldorff, F.B., Nexoe, J., Skovsgaard, C.V., & Olsen, K.R. (2017). Variation in Point-of-Care Testing of HbA1c in Diabetes Care in General Practice. *Int J Environ Res Public Health,* 14.
27. Conti, G., & Ginja, R. (2016). Health Insurance and Child Health: Evidence from Mexico. Discussion paper series pp. 1-21). Bonn, Germany: Institute for the Study of Labor (IZA).
28. Craig, P., Katikireddi, S.V., Leyland, A., & Popham, F. (2017). Natural Experiments: An Overview of Methods,Approaches, and Contributions to Public Health Intervention Research. *Annu Rev Public Health,* 38, 39-56.
29. Wooldridge, J.M. (2002). Econometric analysis of cross section and panel data MIT press. *Cambridge, MA,*108. Craig, P., Katikireddi, S.V., Leyland, A., & Popham, F. (2017). Natural Experiments: An Overview of Methods,Approaches, and Contributions to Public Health Intervention Research. Annu Rev Public Health, 38, 39-56.
30. Gao J, Moran E, Li YF, Almenoff PL.Predicting potentially avoidable hospitalizations. Med Care. 2014,52:164-171.
31. Goble, J.A., & Rocafort, P.T. (2015). Point-of-Care Testing: Future of Chronic Disease State Management? *J Pharm Pract.*

Når lungebetændelse rammer i marts: Monitorering af antibiotikaforbrug

Jens Thusgaard Hørlück, chefkonsulent og faglig leder, DEFACTUM, Region Midtjylland

Motivation

Et af vor tids store problemer er overforbrug af antibiotika. Da forbruget fører til en stadig stigende antal resistente bakterier. Et af de kritiske præparater er det såkaldte pip/tazo (Pipercillin kombineret med Tazobactam og kendt under brandnavnet Tazozin). Pip/tazo skal primært bruges til blodforgiftninger og svære lungebetændelser, men det er et effektivt præparat, som i mange situationer er et sikkert og godt valg, hvis man ser kortsigtet på den enkelte patients interesse. Lægen er derfor i et krydspres mellem patientens interesse, faglige normer og de langsigtede samfundsinteresser.

For at kunne styre forbruget af pip/tazo har man brug for at kunne overvåge eget forbrug og se på en måde, hvor man kan relatere sig til forbruget samtidig med, at man kan se effekter af interventioner.

Udfordringerne er for det første, at ønsket har været en fremgangsmåde, der kan anvendes ned på ugeniveau og på den enkelte enhed. For det andet er det en udfordring, at forbruget er præget af en vis sæsonvariation, primært fordi at svære lungebetændelser svinger henover året og typisk peaker i marts.

Formål

Formålet med projektet er at monitorere forbruget af pip/tazo over tid på en simpel måde, der giver klinikere mulighed for at se, om der er sket ikke-forventede skift i forbruget. I praksis er den første test, om det kan bruges til at belyse effekten af intervention i Region Midtjylland.

Metode

Statistisk proceskontrol (SPC) bliver stadig mere udbredt i sundhedsvæsenet, som en del af en samlet filosofi omkring løbende kvalitetsforbedringer. Et af de mest brugte redskaber fra den statiske proces er kontroldiagrammerne (Shewhart charts, control charts etc.). Med et kontroldiagram kan man følge en proces over tid og skelne mellem tilfældige udsving (*commen cause variation*) eller om der er signal/ikke tilfældig variation (*assignable cause variation*) (7: s 137). I det første tilfælde tolkes det som en stabil proces, og det antages at udsvingene kan tilskrives endogene faktorer. Det tolkes som en ikke stabil proces, at eksogene faktorer har skabt et skift i processen.

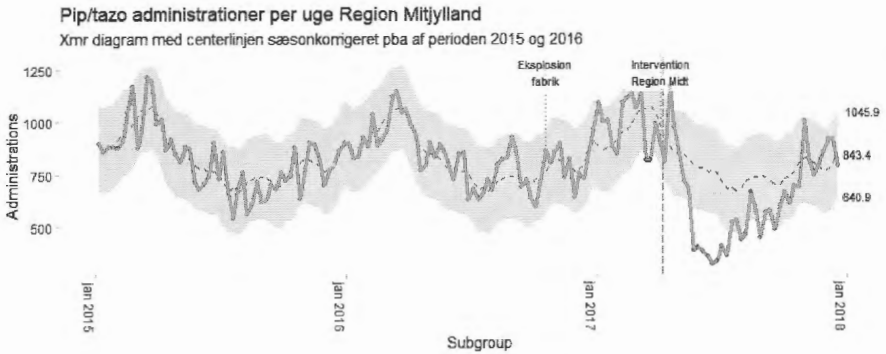
Den måde man typisk har arbejdet på, er at man har indsamlet data fra en baseline periode. Hvis den udviser stabilitet (*common couse variation*), så kan man fastlægge centerlinje samt øvre og nedre grænse og benytte disse fremadrettet.
I et kontroldiagram præsenteres data over tid i et kurvediagram med en række hjælpelinjer, som i kombination med en række regler kan bruges til at afkode, om der er *common cause variation* eller *special cause variation*. For det første er der en såkaldt centerlinje, som typisk er helt flad. Linjen beskriver en central tendens i det viste data, typisk gennemsnittet, medianen eller lignende. For det andet har kontroldiagrammet en øvre og en nedre grænse, som afgrænser det område, hvor man vil forvente at værdierne falder inden for, såfremt processen er præget af tilfældig variation. Det estimeres, at hvis data er normalt fordelt, så er sandsynligheden, for at en given observation falder mere end 3 sigma fra centerlinjen, 1 ud af 370 (7: 2 137). Denne regel er den vigtigste, men der findes en række andre regler som kan identificere *special cause variation.*
Der findes en lang række kontroldiagrammer. Hvilket man skal bruge afhænger af, hvilket data man arbejder med. I dette tilfælde, hvor vi kigger på antal administrationer inden for en given tidsperiode, bruges det såkaldte XmR-diagram(3). I dette er centerlinjen gennemsnittet og grænserne er udregnet på baggrund af gennemsnit af moving range (gennemsnittet plus og minus 2,66 gange gennemsnit af moving range). XmR-diagrammet er udviklet til kontinuerte data. Antallet af administrationer per uge er så højt, at det betragtes som kontinuerte data. XmR-diagrammet har yderligere de fordele, at det er robust overfor brud på antagelsen om normalfordeling, og det kan håndtere moderate niveauer af autokorrelation. XmR-diagrammet dækker over to diagrammet. X-diagrammet viser udviklingen i de observerede værdier over tid og mr-diagrammet viser moving range for den givne periode, heri vises kun x-diagrammet, da mr-diagrammerne ikke viser andet tegn på *special cause variation* på noget tidspunkt.
Den store udfordring er som sagt, at der er en anseelig sæsonvariation, og hvis ikke denne håndteres vil kontroldiagrammet ikke hjælpe os, da det ikke skelner sæsonvariation fra andre eksogene faktorer. Typisk vil man håndtere dette ved enten at tage så lang en periode, at sæsonvariationen udligner sig selv (2). I dette tilfælde vil det kræve et helt års data, hvilket ikke er relevant, da ønsket er at kunne følge udviklingen i realtid eller så tæt på som praktisk muligt. En anden tilgang er at håndtere sæsonvariationen med andre statistiske modeller og så plotte residualerne i kontroldiagrammet (7). Problemet med denne tilgang er, at de plottede data ikke længere er de reelle værdier, hvilket gør det sværere at relatere det til den praksis, man ønsker at monitorere. I stedet vælger vi her at indbygge sæsonkorrektionen i

centerlinjen, så den afspejler det forventede forbrug i en given uge fremfor gennemsnittet. Den øvre og nedre grænse er udregnet på baggrund af centerlinjen, så disse vil også følge sæsonforventningen.
Det forventede forbrug i en given uge estimeres på baggrund af gennemsnittet af forbruget i det samme ugenummer i baselineperioden. Baselineperioden er dog begrænset til to år, og det forventede forbrug bliver meget volatilt, da de ugentlige udsving er ret store. Derfor er det stabiliseret ved først at tage gennemsnittet af de to uger (fra hvert sit år) og igen tage det løbende gennemsnit af det (med 5 ugers hukommelse).
For at få en pejling på fremgangsmådens effektivitet, vil vi se, om den er i stand til at identificere stabilitet (*common cause variation*) i en periode, hvor vi ved der ikke var nogen eksogene påvirkninger og identificere et *special cause variation* i en periode, hvor vi ved der var en udefrakommende påvirkning.

Data

Det anvendte data er administrationer af pip/tazo i Region Midtjylland i perioden fra starten af 2015 til og med 2017. Årene 2015 og 2016 betragtes som relativt stabile år, uden substantielle eksogene påvirkninger som epidemier eller lignende. Året 2017 er anderledes. I første halvdel af 2017 var der en global mangel på pip/tazo på grund af en eksplosion på en fabrik i Kina (1). Region Midtjylland løb ikke tør, men der blev iværksat en intervention, som skulle få klinikere til at flytte forbruget af pip/tazo over på andre præparater.

Resultater



Resultaterne vises i ovenstående graf. Helt overordnet så viser diagrammet set *common cause* variation i baselineperioden (1. jan 2015 til 31. dec. 2016) hvilket indikerer det en vis stabilitet/forudsigelighed. I starten af 2017 er der nogle enkelte

uger der ikke falder uden for grænserne – både over og under. Efter interventionen ses et markant fald efterfulgt af næsten 4 måneder hvor antallet af administrationer ligger under den nedre grænse, mens antallet af administrationer gradvist bevæger sig tilbage til det forventede område.

Diskussion

Der er flere aspekter, der kan diskuteres. For det første kan man overveje, om selve sæsonkorrektionen kan gøres bedre, om den reelt er overfittet til 2015/2016 data, og hvilken værdi det har, at 2015 og 2016 falder inden for en det forventede område. Det kunne være interessant, at basere sæsonkorrektionen på en langt længere periode, men data findes desværre kun i perioden fra 2015 og frem.

I teorien kunne man dog gå helt anderledes til det, og i stedet for sæsonkorrektion, justere forbruget for en række faktorer der har indflydelse på behovet for pip/tazo, og så i stedet kigge på det justerede forbrug. I det tilfælde ville man så umiddelbart ikke have fordelen, at man kigger på det reelle forbrug.

For at XmR diagrammer med sæsonkorrigeret centerlinje, kan blive et rigtig brugbart redskab til at reducere forbruget af antibiotika, så kræver det, at man kan kigge på en bestemt afdelings data. På nuværende tidspunkt kræver det nærmere undersøgelse, at belyse om sæsonkorrektionen giver mening på afdelingsniveau.

Referencer

1. Anonymous. Pollutants under alert levels after East China pharmaceutical factory. China Daily. 2016.
2. Wheeler DJ. Making Sense of Data: SPC for the Service Sector. Knoxville, TN: SPC Press, 2003.
3. Wheeler DJ. Understanding Variation: The Key to Managing Chaos, 2nd Edition. Knoxville, TN: SPC Press, 2000.
4. Anhoj J. Diagnostic value of run chart analysis: using likelihood ratios to compare run chart rules on simulated data series. PloS one 2015; 10(3): e0121349.
5. R_Core_Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2017.
6. Anhoj J, Wentzel-Larsen T. Sense and sensibility: on the diagnostic value of control chart rules for detection of shifts in time series data. BMC medical research methodology 2018; 18(1):

7. Mohammed MA1, Worthington P, Woodall WH. Plotting basic control charts: tutorial notes for healthcare practitioners.Qual Saf Health Care. 2008 Apr;17(2):137-45. doi: 10.1136/qshc.2004.012047.

Ranking systems - a discussion

Gorm Gabrielsen, Associate Professor Emeritus, CBS
G.G. Consulting Aps

Ranking system has become extremely popular in the social sciences. In almost all areas of public life one must knew who are best, largest, earns most money, have the highest reputation etc. etc.

Construction of a ranking consists of comparing objects. In most cases, this comparison is performed indirectly. In practice (a lot of) aspects of each object is collected or constructed and combined into a single "measurement" of each object. These measurements are considered imbedded into some kind of (external) measurement scale. From the specific measurements, the objects are compared using the external measurement scale.

In general very little is said about the characteristics or properties of such a scale. Is there a metric or is it just an order scale? If there is a metric what is the structure of the metric? In general, these kind of questions are swept of by declaring the measuring practice as a procedure.

There are however sometimes also the possibility of direct comparisons. This is often used in sports games. We will as an example consider handball tournaments. A tournament can be performed in different ways.

In a usual tournament, each team plays against all the other teams. The result of a match is recorded and transformed into some points. At the end of the tournament the points of a team is summed and these sums is used to rank-order the teams.

In a CUP- tournament, the outcome of a match is determining for which of the two teams is excluded from the rest of the cup-tournament.

Making a metric rating scale for the teams in VM 2019

Handball for men

We need to have a thinking-model to combine the statistical model with "the real world". Meaning how can I explain the result of the match and the embedding into a scale to "sports-people".

Thinking model

An often-used way of considering a match is to split the skills of a team into two separate skills, each team has skills in attack and skills in defense. Let a_i denotes the skill of attack of *Team i* and d_i denotes the skill of defense of *Team i*.

When *Team i* and *Team j* meets, the attack of *Team i* plays against the defense of *Team j* and the attack of *Team j* plays against the defense of *Team i*.

The odds of a success of an attack of *Team i* against the defense of *Team j* can be modelled as a_i/d_j.

Correspondently the odds a success of an attack of *Team j* against the defense of *Team i* can be expressed as a_j/d_i. If $a_i/d_j > 1$ the attack strength of *Team i* is greater than the defense strength of

Team j's defense strength and vice versa if $a_i/d_j < 1$, *Team i's* attack strength is less than Teams j's
defense strength.
If $a_i/d_j = 1$ *Team i*'s attack strength is balanced by *Team j*'s defense strength.
The odds of *Team i* winning when *Team i* meats *Team j* is expressed as $a_i/d_j/a_j/d_j$. (1)
Now $a_i/d_j/a_j/d_j = a_i \times d_i/a_j \times d_j$.
The overall strength of *Team i, denoted* α_i, therefore can be modelled by $a_i \times d_i = \alpha_i$.
The odds of *Team i* winning if *Team i* meats *Team j* is denoted *Odds(i,j)* and becomes
$Odds(i,j) = \alpha_i/\alpha_j$.
We use ofte log Odds defined by
$\log \text{Odds}(i,j) = \log(\alpha_i) - \log(\alpha_j) = \beta_i - \beta_j$, where $\log(\alpha_i) = \beta_i$. The β's are referred to as the additive parameters.

Estimating the Odds or log Odds

When *Team i* meets *Team j* we collect the number of scored goals. We denote;
s1(i,j) is the number of goals scored by *Team i* (the first coordinate),
s2(i,j) is the number of goals scored by *Team j* (the second coordinate).
(With this notation s1(i,j) = s2(j,i)).
In the present paper we have chosen to let the observed log Odds(i,j) be
Observed log Odds(i,j) = [s1(i,j)]/ [s1(i,j) + s2(j,i)] - [s2(j,i)]/ [s1(i,j) + s2(j,i)]
That is the difference between the relative numbers of scored goals.
We can think of this as the observed $(\beta_i - \beta_j)$ or the raw observed $(\beta_i - \beta_j)$.
Note that we do not want to calculate β_i or β_j, but only the difference.
There are other possibilities e.g. $\beta_i - \beta_j \sim$ s1(i,j) - s2(j,i)) but the choice must be tested in practice.
The next question is whether these observed odds can be arranged into a common structure such that
$Odds(i,j) = Odds(i,k) \times Odds(k,j)$ *for all i,j,k.*
This means that the observed Odds can be considered to be imbedded into a ratio scale.
Such that the underlying structure is a ratio scale.

We do this by estimating the structure (model-based inference) and study the properties of the estimated structure among others checking the goodness of fit. In the estimation procedure we uses the additive model.

Data

The 24 teams is divided into 4 groups A, B, C and D, Table 1. The six teams in a group plays a tournament that is every team plays a match against each of the other teams in the group. This gives 15 matches in each group. The three best teams in each group continuous to the CUP tournament.

At first for each of the four groups, we will study whether the six teams in a group can be arranged into a metric scale – and try to give an interpretation of that question.

Table 1 The 24 teams/countries and there groups

Team number	*Country*	*Group*		Team number	*Country*	*Group*
1	Germany	A		13	Chile	C
2	Serbia	A		14	Saudi-Arab	C
3	Brazil	A		15	Tunisia	C
4	Russia	A		16	Austria	C
5	France	A		17	Norway	C
6	Korea	A		18	Denmark	C
7	Japan	B		19	Angola	D
8	Iceland	B		20	Argentina	D
9	Bahrain	B		21	Egypt	D
10	Macedonia	B		22	Qatar	D
11	Croatia	B		23	Hungary	D
12	Spain	B		24	Sweden	D

Analysis of Group A

The results for Group A are shown in Table 2, where the 15 matches are recorded with the teams, the goals of each team are also recoded. Furthermore, we have calculated the difference in goals and the relative difference, which is the observed log Odds.

Table 2 Group A matches

Match	*Country Left*	*Country Right*	*Goals Left*	*Goals Right*	*Goals DIF Right - Left*	*Observed log Odds*	*DIF fitted*	*Cook's distance*
1	Germany	Korea	30	19	-11	-.22	-10.8	0.00
2	Serbia	Russia	30	30	0	.00	4.17	0.16
3	Brazil	France	22	24	2	.04	3.94	0.06
4	Russia	Korea	34	27	-7	-.11	-7.77	0.01
5	Germany	Brazil	34	21	-13	-.24	-6.01	0.54
6	France	Serbia	32	21	-11	-.21	-7.35	0.16
7	Serbia	Brasilien	22	24	2	.04	2.44	0.00
8	Russia	Germany	22	22	0	.00	4.09	0.29
9	France	Korea	34	23	-11	-.19	-11.2	0.00
10	Russia	Brazil	23	25	2	.04	-0.78	0.11
11	Korea	Serbia	29	31	2	.03	3.47	0.02
12	Germany	France	25	25	0	.00	-1.19	0.02
13	Brazil	Korea	35	26	-9	-.15	-6.77	0.04
14	Germany	Serbia	31	23	-8	-.15	-8.77	0.01
15	France	Russia	23	22	-1	-.02	-3.12	0.07

To study whether the observed log Odds can be considered to have an additive structure we apply the model:

$E(\text{Observed Log Odds(team right, team left)}) = \beta_{right} - \beta_{left}$.

Here the "double index" (team right, team left) refers to the 15 observation in table 1.

The model is a general linear model, and we apply OLS estimators. To identify the parameters we use the restriction to the parameters that they are centralized, i.e. the sum is zero, that is

$\sum \beta_{team} = 0$, where the sum is over the 6 teams (countries) in the group.

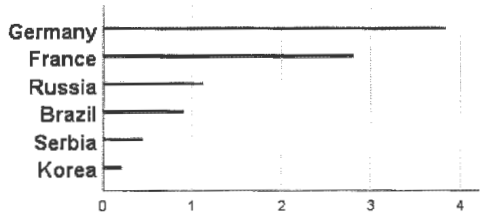
The result is found in Table 2 A. Further, we calculate the standardized estimates by the restriction that the standard deviation of the parameters are 1. Finally, to get back to the *overall strength of Team i, denoted* α_i, we apply the exponential function to the standardized and centralized parameters.

Table 3 Group A Parameter estimates

Group A	country	centralized	standardized	exponential
1	Germany	0.102	1.342	3.827
2	Serbia	-0.061	-0.806	0.447
3	Brazil	-0.008	-0.104	0.901
4	Russia	0.008	0.112	1.119
5	France	0.078	1.027	2.794
6	Korea	-0.119	-1.572	0.208

How can we represent the estimated metric ordering – i.e. the ration scale - or the estimated odd Ratio? We can read it from the last column in Table 3. We can also make an illustration Figure 4 Group A from which we can read how good a team is relative to another team. For example the best team in group A Germany is 1.37 (= 3.827/2.794) times better than the second best team, France, but 3.42 (=3.827/1.119) times better than the third team, Russia. The point "1" at the x-axis is the geometric mean of the strength of the six teams and the scale number referrers to how deviant a team is from the geometric mean. The strength is in this way given at Indices with the base "1". So we can conclude that the two best teams are Germany and France, in the middle we have Russia and Brazil and lowest Serbia and Korea.

Figure 4, Group A Team Strength



How good is the model to describe the observed log Odds? We calculate the usual goodness-of fit as 0.206/0.251 = 0.822. We can think of this as R^2 (although there is no constant in the model). This tells that 82.2 % of the total variation in the 15 observations of log Odds can be ascribed to an existing underlying structure of a metric ordering scale of the 6 teams (countries). The remaining 17.8% is due to random variations. Whether this is much depends on the use of the present analysis.

We can also ask in another way whether there is anything systematic in the ordering of the 6 teams. We can clarify this by testing a hypothesis of no difference in the strength of the 6 teams, which gives $F(5,10) = 9.216$, $p=0.002$, concluding that we have some systematic structure.

Further, to study the origin of the residual17.8 % of variation we calculated cook' distance for each of the 15 observations. As Cook's distance measures the effect of deleting a single given observation, we can study whether there is a single outcome of a match, which is outlying. Cook's distances is included in Table 2. It is match 5 between Germany and Brazil and to some degree match 8 between Russia and Germany which courses trouble as is seems that the strength of Germany is not consistent during the games.
If Match 5 is removed we get $R^2 = 89.5\%$, confirming that the winning of Germany in Match 5 "is unexpected" high.
One can also note that if match 5 is removed *Team 3* and *Tem 4* is changing the order such that Brazil will be included in the next "middle-play" instead of Russia.

Analysis of Group B
The results for Group B are shown in Table 5, where the 15 matches are recorded with the teams, the goals of each team are recoded. Furthermore, we have calculated the difference in goals and the relative difference, which is the observed log Odds.

Figure 5 Group B Matches

Match	*Country Left*	*Country Right*	*Goals Left*	*Goals Right*	*Goal DIF Right - Left*	*Observed log Odds*	*DIF fitted*	*Cook' distonce*
16	Japan	Makedonia	29	38	9	,13	2.99	0.17
17	Iceland	Croatia	27	31	4	,07	4.08	0.00
18	Bahrain	Spain	23	33	10	,18	12.28	0.03
19	Makedonia	Bahrain	28	23	-5	-,10	-5.36	0.00
20	Croatia	Japan	35	27	-8	-,13	-11.88	0.08
21	Spain	Iceland	32	25	-7	-,12	-2.14	0.15
22	Iceland	Bahrain	36	18	-18	-,33	-9.81	0.48
23	Croatia	Makedonia	31	22	-9	-,17	-7.79	0.01
24	Spain	Japan	26	22	-4	-,08	-7.62	0.12
25	Japan	Iceland	21	25	4	,09	5.58	0.02
26	Croatia	Bahrain	32	20	-12	-,23	-13.11	0.01
27	Makedonia	Spain	21	32	11	,21	6.06	0.18
28	Bahrain	Japan	23	22	-1	-,02	2.72	0.14
29	Makedonia	Iceland	22	24	2	,04	3.53	0.02
30	Spain	Croatia	19	23	4	,10	1.38	0.08

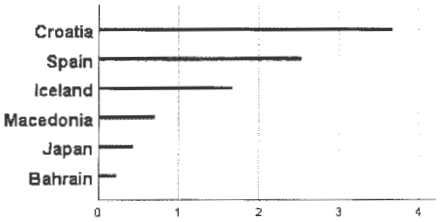
The result of the parameter estimations is found in Table 6.

Table 6 Group B Parameter estimates

Group B	country	centralized	standardized	exponential
7	Japan	-0.076	-0.853	0.426
8	Iceland	0.045	0.509	1.663
9	Bahrain	-0.136	-1.531	0.216
10	Macedonia	-0.031	-0.352	0.703
11	Croatia	0.116	1.297	3.660
12	Spain	0.083	0.929	2.533

The strength parameters is shown in Figure7 from which we can read how good a team is relative to another team. For example the best team at group A Croatia is 1.44 (= 3.660/2.533) times better than the second best team, Spain. The point "1" at the x-axis is the geometric mean of the strength of the six teams and the scale number referrers to how deviant a team is from the geometric mean. The strength is in this way given at Indices with the base "1". So we can conclude that the best team is Croatia and the strength of the other teams is decreasing softly.

Figure 7 Group B Team Strength



The goodness-of fit is $R^2 = 79.8\%$ which this tells that 79.8 % of the total variation in the 15 observations of log Odds can be ascribed to an existing underlying structure of a metric ordering scale of the 6 teams (countries). The remaining 20.2% is due to random variations. Whether this is much depends on the use of the present analysis.
We can also ask in another way whether there is anything systematic in the ordering of the 6 teams. We can clarify this by testing a hypothesis of no difference in the strength of the 6 teams, which gives $F(5,10) = 7.907$, $p=0.003$, concluding that we have some systematic structure.
Further, Cook's distance measures the effect of deleting a single given observation; we can study whether there is a single outcome of a match, which is outlying. Cook's distances is included in Table 3. It is match 22 between Island and Bahrain and to some degree match 8 between Makedonia and Spain and maybe Bahrain and Japan that the strength of Bahrain is not consistent during the games.
If Match 22 is removed we get $R^2 = 86.18\%$.

Analysis of Group C
The results for Group C is shown in Table 8, where the 15 matches are recorded with the teams, the goals of each team are recoded. Furthermore, we have calculated the difference in goals and the relative difference, which is the observed log Odds.

Table 8 Group C Matches

Match	Country Left	Country Right	Gools Left	Gools Right	Goals DIF Right - Left	Observed log Odds	DIF fitted	Cook' distance
31	Chile	Denmark	16	39	23	,42	16.64	0.25
32	Saudi-Arab	Austria	22	29	7	,14	2.82	0.13
33	Tunisia	Norway	24	34	10	,17	10.4	0.00
34	Austria	Chile	24	32	8	,14	-1.11	0.49
35	Norway	Saudi-Arab	40	21	-19	-,31	-17.97	0.01
36	Denmark	Tunisia	36	22	-14	-,24	-12.92	0.01
37	Tunisia	Chile	36	30	-6	-,09	-5.26	0.00
38	Norway	Austria	34	24	-10	-,17	-13.88	0.08
39	Denmark	Saudi-Arab	34	22	-12	-,21	-18.93	0.29
40	Saudi-Arab	Tunisia	20	24	4	,09	5.07	0.01
41	Norway	Chile	41	20	-21	-,34	-15.81	0.14
42	Austria	Denmark	17	28	11	,24	12.73	0.03
43	Chile	Saudi-Arab	32	27	-5	-,08	-2.09	0.05
44	Austria	Tunisia	27	32	5	,08	3.54	0.01
45	Denmark	Norway	30	26	-4	-,07	-2.43	0.01

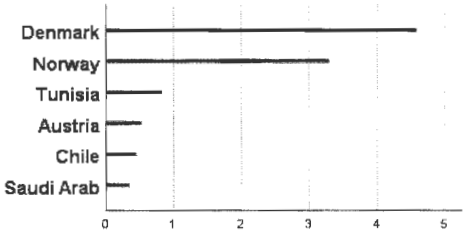
The result of the parameter estimations for Group C is found in Table 9.

Table 9 Group C Parameter estimates

Group C	country	centralized	standardized	exponential
13	Chile	-0.104	-0.802	0.448
14	Saudi-Arab	-0.140	-1.075	0.341
15	Tunisia	-0.025	-0.189	0.828
16	Austria	-0.085	-0.65	0.522
17	Norway	0.155	1.191	3.290
18	Denmark	0.198	1.525	4.594

The strength parameters is shown in Figure 10 from which we can read how good a team is relative to another team. The best team at group C is Denmark which is 1.40 (= 4.594/3.290) times better than the second best team, Norway. The point "1" at the x-axis is the geometric mean of the strength of the six teams and the scale number referrers to how deviant a team is from the geometric mean. The strength is in this way given as indices with the base "1". Therefore, we can conclude that the best team is Denmark and Norway being much strengthen than the remaining 4 teams. Further performing a test of Denmark and Norway being of equal strength gives p = 0.421 – but there are very few degrees of freedom. The remaining 4 teams are significantly weaker compared to Denmark and Norway

Figure 10 Group C Team Strength



The goodness-of fit in Group D is $R^2 = 88.3\%$ which this tells that 88.3.8 % of the total variation in the 15 observations of log Odds can be ascribed to an existing underlying structure of a metric ordering scale of the 6 teams (countries).
We can also ask in another way whether there is anything systematic in the ordering of the 6 teams. We can clarify this by testing a hypothesis of no difference in the strength of the 6 teams, which gives $F(5,10) = 15.167$, $p=0.000$, concluding that we have some systematic structure.

With so high a R^2 we would not expect to find any interesting using Cook's distance as every small deviance would be reflected in a high Cook's distance.

Analysis of Group D
The results for Group D is shown in Table 11, where the 15 matches are recorded with the teams, the goals of each team are recoded. Furthermore, we have calculated the difference in goals and the relative difference, which is the observed log Odds.

Table 11 Group D Matches

Match	*Country Left*	*Country Right*	*Goals Left*	*Goals Right*	*Goals DIF Right - Left*	*Observed log Odds*	*DIF fitted*	*Cook' distance*
46	Angola	Qatar	24	23	-1	-.02	5.04	0.27
47	Argentina	Hungary	25	25	0	.00	4.13	0.11
48	Egypt	Sweden	24	27	3	.06	6.77	0.09
49	Qatar	Egypt	28	23	-5	-.10	0.02	0.16
50	Hungary	Angola	34	24	-10	-.17	-8.71	0.01
51	Sweden	Argentina	31	16	-15	-.32	-8.12	0.36
52	Hungary	Qatar	32	26	-6	-.10	-2.49	0.06
53	Argentina	Egypt	20	22	2	.05	1.68	0.00
54	Sweden	Angola	37	19	-18	-.32	-13.46	0.11
55	Angola	Argentina	26	33	7	.12	3.99	0.04
56	Hungary	Egypt	30	30	0	.00	-2.56	0.03
57	Qatar	Sweden	22	23	1	.02	5.99	0.20
58	Egypt	Angola	33	28	-5	-.08	-6.57	0.01
59	Qatar	Argentina	26	25	-1	-.02	-2.03	0.01
60	Sweden	Hungary	33	30	-3	-.05	-5.68	0.03

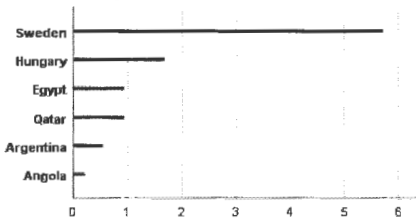
The result of the parameter estimations for group D is found in Table 12.

Table 12 Group D Parameter Estimates

Group B	country	centralized	standardized	exponential
19	Angola	-0.112	-1.524	0.218
20	Argentina	-0.045	-0.606	0.545
21	Egypt	-0.005	-0.062	0.94
22	Qatar	-0.005	-0.066	0.936
23	Hungary	0.038	0.517	1.677
24	Sweden	0.128	1.742	5.708

The strength parameters is shown in Figure 10 from which we can read how good a team is relative to another team. The best team in Group C is Sweden which is 3.40 (= 5.708/1.677) times better than the second best team, Hungary. The point "1" at the x-axis is the geometric mean of the strength of the six teams and the scale number referrers to how deviant a team is from the geometric mean. The strength is in this way given as indices with the base "1". So, we can conclude that the best team is Sweden being much strengthen than the remaining 5 teams.

Figure 13 Group D Team Strength



The goodness-of fit in Group D is $R^2 = 0.684$ which this tells that 68.4 % of the total variation in the 15 observations of log Odds can be ascribed to an existing underlying structure of a metric ordering scale of the 6 teams (countries). The remaining 31.6% is due to random variations.

We can also ask in another way whether there is anything systematic in the ordering of the 6 teams. We can clarify this by testing a hypothesis of no difference in the strength of the 6 teams, which gives $F(5,10) = 4.321$, $p=0.024$, concluding that we have some systematic structure but *very* much random variation.

The largest Cook's distance is from the match 51 between Sweden and Argentina where Sweden a shows high strength. However, the next matches with high Cook's all involve Qatar. It seem that Qatar has a very inconsistent way of playing. It looks like the strength of the *Team Qatar* depends on which team they play against.
If the 5 matches involving Qatar is removed we get $R^2 = 85.1$, indicating that the misfit is due to Qatar.
This means that the thinking model breaks down.

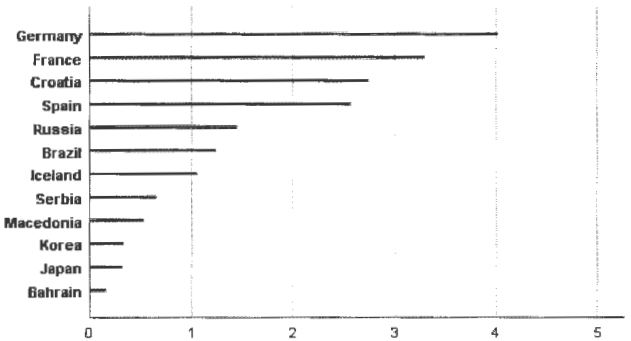
Conclusion of the analyses of the four groups
It seems that in most cases it works to imbed the ordering into a metric ratio scale, but:
For some matches the result are to some degree unexpected, but that is maybe not so strange.
Some teams are acting in an inconsistent way, Qatar, such that the strength of the team depends of the team they are meeting. How should that be understood?

In the next middle play the teams in Group A and Group B meets.
And the teams in Group C and D meats.

Group A and Group B
The common estimated strength is shown in Figure 14.
Note that the ordering of the teams in Group A is maintained and similarly the ordering of the teams in Group B is maintained. The two orderings are intertwined. In addition, the relative strength of the teams from each group is approximately maintained.

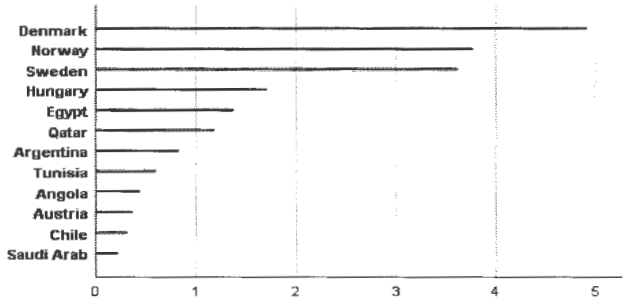
The common R^2 becomes $R^2 = 75.6$ % which is not so much. It the same probles as seen in Group A and Group B.

Figure 14 Group A and Group B Team Strength



Group C and Group D
The common estimated strength is shown in Figure 15.
Note that the ordering of the teams in Group A is maintained and similarly the ordering of the teams in Group B is maintained. The two orderings are intertwined. In addition, the relative strength of the teams from each group is approximately maintained.

Figure 15 Group A and Group B Team Strength



The common R^2 becomes $R^2 = 82.9$ % which is very nice.

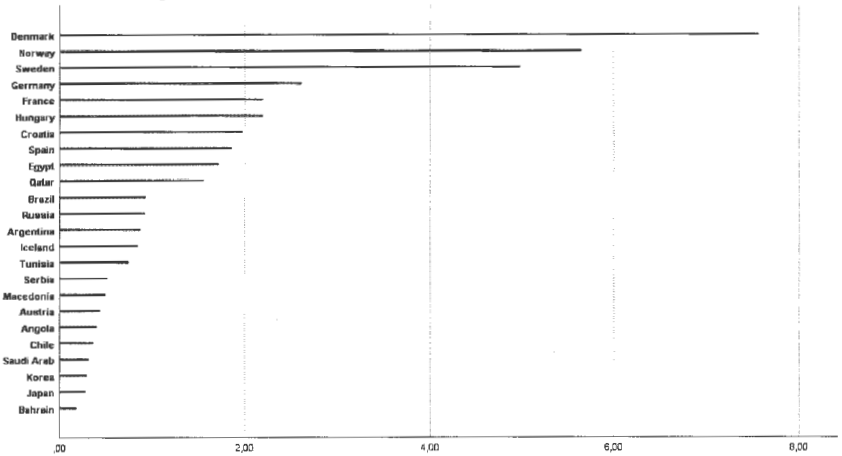
All matches leaving out finale and semi-finale gives the metric rank-order Table 16 and Figure 17.
So, we can use this to make bettings for the finale.

The odds e.g. for the Final match Denmark - Norway becomes 7.57/5.64 = 1.34.

Table 16, Total metric ordering

7.57	Denmark
5.64	Norway
4.98	Sweden
2.61	Germany
2.19	France
2.19	Hungary
1.97	Croatia
1.86	Spain
1.71	Egypt
1.55	Qatar
0.93	Brazil
0.92	Russia
0.87	Argentina
0.85	Iceland
0.75	Tunisia
0.51	Serbia
0.5	Macedonia
0.44	Austria
0.4	Angola
0.37	Chile
0.32	Saudi Arab
0.3	Korea
0.28	Japan
0.18	Bahrain

Figur 17 Team Strength



Likelihoodprincippet og den klassiske p-værdi

Tom Engsted[1]
Institut for Økonomi, Aarhus Universitet
Fuglesangs alle 4, 8210 Aarhus V.
Email: tengsted@econ.au.dk

December 2019

Abstract: *Ifølge likelihoodprincippet (LP) indeholder likelihoodfunktionen al relevant information fra stikprøven om en given models parametre, og likelihoodfunktioner, der er proportionale med hinanden, fører til samme statistiske inferens. Men klassiske frekvensbaserede hypotesetests overholder ikke LP. Det diskuteres, om statistiske analyser indenfor samfundsvidenskab bør overholde LP og - i så fald - hvilke metoder, man da alternativt kan anvende.*

Keywords: *Frekvensbaserede vs. bayesianske metoder, eksperimentelle setup, stop-regel princippet, samfundsvidenskabelig praksis.*

1. Indledning.

De senere år har der været et stigende fokus på det velkendte problem, at man indenfor flere empiriske videnskaber har svært ved at reproducere tidligere resultater. Man taler om *the reproducibility crisis*, *false discoveries*, *p-hacking*, etc., og diskuterer nu indgående, om de statistiske metoder, vi traditionelt anvender i empiriske analyser, er en medvirkende årsag til problemet, se eksempelvis Wasserstein and Lazar (2016), Harvey (2017), McShane et al. (2019) og Wasserstein et al. (2019).

Anvendelse af klassiske frekvensbaserede statistiske metoder har været den fremherskende tilgang indenfor samfundsvidenskab. Ikke mindst de klassiske hypotesetests spiller en vigtig rolle i denne tilgang. I sidste års symposieindlæg (Engsted, 2019) sammenlignede jeg sådanne tests med bayesianske

[1]Tak til Mikkel Bennedsen og Carsten Tanggaard for deres villighed til i årets løb at diskutere statistisk metodologi med mig. De er naturligvis ikke ansvarlige for fejl og mangler i denne artikel, ligesom de heller ikke nødvendigvis er enige i artiklens budskaber.

hypotesetests og illustrerede, hvordan de klassiske tests for et givet datasæt generelt giver en væsentlig stærkere indikation af 'signifikans', end de bayesianske tests. Hvis de tests vi sædvanligvis anvender har en tendens til for ofte at finde sammenhænge eller effekter, kan det være en del af forklaringen på den manglende reproducerbarhed og de mange *false discoveries*.

I nærværende artikel diskuterer jeg et relateret problem, nemlig at vi i samfundsvidenskabelige empiriske analyser oftest ikke har kontrol med - eller kender - det præcise eksperimentelle setup, der ligger til grund for de data vi arbejder med, og hvor eksakt gentagelse af eksperimentet er enten umuligt eller svært at forestille sig. Jeg vil argumentere for, at dette udgør er særligt problem for de klassiske frekvensbaserede tests, da disse forudsætter en kendt og fuldt specificeret stikprøveplan, hvor tænkte gentagelser af eksperimentet giver mening. Den klassiske p-værdi angiver sandsynligheden for de observerede data eller mere ekstreme data under nulhypotesen, H_o, og er dermed baseret på hypotetiske data, der ikke er observeret, men som *kunne* være observeret (i tænkte gentagelser af eksperimentet) hvis H_o er sand. Disse hypotetiske data afhænger afgørende af det præcise eksperimentelle setup (Wagenmakers, 2007). Den klassiske p-værdis afhængighed af hypotetiske data i tænkte gentagelser indebærer, at p-værdien ikke overholder *likelihoodprincippet* (LP).

Metoder, der overholder LP, derimod - eksempelvis visse bayesianske metoder -, bygger ikke på tænkte gentagelser af eksperimentet og er dermed ikke på samme måde følsomme overfor det præcise eksperimentelle setup (Berger and Wolpert, 1988). Jeg illustrerer disse pointer med konkrete eksempler, og diskuterer implikationer for videnskabelig praksis indenfor empirisk samfundsvidenskab.

2. Klassiske tests afhængighed af det eksperimentelle setup.

De klassiske frekvensbaserede metoder bygger på '*the principle of repeated sampling*', hvor "*statistical procedures are to be assessed by their behaviour in hypothetical repetitions under the same conditions. ... Measures of uncertainty are to be interpreted as hypothetical frequencies in long run repetitions.*" (Cox and Hinkley, 1974, p.45). Denne tilgang forudsætter en ret detaljeret og præspecificeret stikprøveplan, hvor man eksempelvis på forhånd præcist fastlægger stikprøvestørrelsen, dvs. antal observationer i stikprøven (se eksemplet i afsnit 2.1 nedenfor).

Den klassiske tilgang er velegnet i situationer, hvor man har kendskab til - og fuld kontrol over - det eksperimentelle setup, hvor man rent faktisk er i stand til at gentage eksperimentet, og hvor kontrol over fejlsandsynlighederne (Type-1 og Type-2 fejl) i sig selv er et vigtigt mål med analysen (Neyman and Pearson, 1933). Eksempelvis statistisk kvalitetskontrol i produktionsvirksomheder, hvor man ønsker at minimere andelen af fejlbehæftede 'dimsedutter' under hensyntagen til, at der er et tradeoff forbundet med på den ene side at have en meget præcis men omkostningstung kvalitetskontrol, og på den anden side at lade fejlbehæftede 'dimsedutter' passere en omkostningslet men uprækis kvalitetskontrol.

Ovenstående ideelle situation udspiller sig dog sjældent i praksis. Samfundsforskere, der anvender de klassiske metoder, tilvejebringer og analyserer ofte deres datamateriale på måder, der langt fra er i overensstemmelse med en fuldt specificeret og forudbestemt stikprøveplan. Eksempelvis i opinionsundersøgelser, hvor det præcise antal respondenter som oftest ikke er kendt på forhånd, men hvor data alligevel analyseres som om stikprøvestørrelsen er forudbestemt. Dataindsamlingen kan være begrænset til en given periode, hvor man ikke ved, hvor stor stikprøve man har når perioden udløber; alligevel tester man som om stikprøvestørrelsen er kendt og specificeret på forhånd. I visse eksperimenter analyserer man data sekventielt som de indløber undervejs, men uden at tage hensyn til, at stikprøvestørrelsen dermed ikke er forudbestemt, og uden at korrigere signifikansniveauet for, at der foretages multiple tests. Ofte ligger der slet ikke et eksperimentelt setup til grund for stikprøven; data er i stedet passivt observerede resultater af udviklingen i samfundet (BNP, inflation, aktiekurser, etc.), og hvor det er svært at forestille sig 'eksperimentet' gentaget. At det kan gøre en afgørende forskel for den statistiske analyse, om data er eksperimentelle eller ikke-eksperimentelle, har været kendt længe (se eksempelvis Leamer, 1978), men det er ikke altid at man i empiriske analyser er opmærksom på denne forskel.

Vigtigheden af princippet om *repeated sampling* kan illustreres med det velkendte klassiske 95% konfidensinterval. Et sådant interval, baseret på en given stikprøve, fortolker vi ofte uformelt som at den sande populationsparameter med 95% sandsynlighed ligger indenfor det beregnede interval. Men da populationsparametre i klassisk statistik ikke er stokastiske, er denne fortolkning ikke korrekt. For et givet interval ligger parameteren indenfor intervallet med sandsynlighed enten 0 eller 1. Den korrekte fortolkning er i stedet: hvis man udtrækker mange stikprøver fra populationen, og beregner et 95% konfidensinterval i hver af disse, vil 95% af disse intervaller indeholde

den sande populationsparameter.

Den klasiske p-værdis afhængighed af det præcise eksperimentelle setup og af hypotetiske data i tænkte gentagelser af eksperimentet, illustreres i nedenstående eksempel.

2.1 Eksempel: Binomial vs. negativ-binomial setup.
Et velkendt eksempel til illustration af problemstillingen er følgende (jf. Berger and Wolpert, 1988, Example 9; Wagenmakers, 2007, Example 4): Antag at der er udført $n = 12$ uafhængige bernoulliforsøg, hver med sandsynlighedsparameter θ, og hvor resultatet er enten 'succes' (1) eller 'fiasko' (0) i hvert forsøg. Eksperimentet resulterer i følgende stikprøve:

Forsøg nr.	1	2	3	4	5	6	7	8	9	10	11	12
Resultat	1	1	0	1	1	1	0	1	1	1	1	0

Hvis stikprøvestørrelsen på $n = 12$ er kendt og fastlagt på forhånd, kan antal succes i dette eksperiment beskrives ved en binomialfordeling. Lad X være antal succes i n forsøg; da fås sandsynligheden for $X = s$ som:

$$P(X = s \mid n, \theta) = \binom{n}{s} \theta^s (1-\theta)^{n-s}. \tag{1}$$

Vi ønsker at teste hypotesen H_o: $\theta = 0.5$ overfor et enkeltsidet alternativ H_1: $\theta > 0.5$. Det klassiske eksakte binomialfordelingstest giver da følgende p-værdi:

$$\text{p-værdi} = \sum_{s=9}^{12} \binom{12}{s} (0.5)^s (1-0.5)^{12-s} = 0.073. \tag{2}$$

Med en p-værdi på 0.073 kan vi ikke forkaste H_o på det sædvanlige 5% signifikansniveau.

Antag nu i stedet, at det eksperimentelle setup er at gentage bernoulliforsøget indtil der er fremkommet 3 fiaskoer, og at dette resulterer i netop ovenstående stikprøve. Hermed er den underliggende fordeling ikke længere binomialfordelingen, men den *negative binomialfordeling*, der kan beskrives som følger: Lad Y være antal forsøg indtil $f = 3$ fiaskoer er opnået; da fås sandsynligheden for $Y = n$ som:

$$P(Y = n \mid f, \theta) = \binom{n-1}{f-1} \theta^s (1-\theta)^{n-s}. \quad (3)$$

Med samme hypotese som ovenfor, fås p-værdien nu som

$$\text{p-værdi} = \sum_{n=12}^{\infty} \binom{n-1}{2} (0.5)^s (1-0.5)^{n-s} = 0.033. \quad (4)$$

Bemærk hvordan p-værdien er forskellig i de to tilfælde. I binomialforsøget forkastes H_o ikke på et 5% signifikansniveau, mens hypotesen forkastes i det negative binomialforsøg. *Dette på trods af, at stikprøven i begge tilfælde er eksakt den samme og baseret på uafhængige bernoulliforsøg med samme parameter* θ. Forklaringen er de ikke-observerede hypotetiske værdier under H_o, der er forskellige i de to forsøg.

Eksemplet illustrerer hvordan man i den klassiske frekvensbaserede tilgang er nødt til at kende den præcise stikprøveplan for at kunne foretage korrekt statistisk inferens. I næste afsnit diskuteres dette nærmere, og der gennemgås en alternativ tilgang, der ikke stiller lige så strenge krav til kendskab til det præcise eksperimentelle setup.

3. Likelihoodprincippet og dets implikationer.

Simpelt udtrykt siger *likelihoodprincippet* (LP), at likelihoodfunktionen indeholder al relevant information fra stikprøven om en given models parametre, og at likelihoodfunktioner, der er proportionale med hinanden, skal føre til samme statistiske inferens. LP indebærer, at kun den faktisk observerede stikprøve er relevant for inferensen. Berger and Wolpert (1988) giver en detaljeret beskrivelse af forskellige udgaver af LP og dets matematiske grundlag, samt implikationer og anvendelser af princippet. I nærværende artikel begrænser jeg mig til at diskutere LP i relation til hypotesetestning og almindelig statistisk praksis indenfor samfundsvidenskab.

3.1 Frekvensbaserede og bayesianske hypotesetests.

Eksemplet i afsnit 2.1 med et binomial vs. negativ-binomial setup illustrerer, hvordan det klassiske frekvensbaserede test er i modstrid med LP: Likelihoodfunktionen i de to eksperimenter er givet ved henholdsvis (1) og (3). Det ses, at disse to funktioner er proportionale med hinanden, da den del,

der afhænger af θ, er identisk ($\theta^s(1-\theta)^{n-s}$). Ifl. LP bør inferensen fra de to eksperimenter derfor være identisk, men det er ikke tilfældet, da de to p-værdier er forskellige.

Årsagen til at den klassiske p-værdi er i modstrid med LP, er, at p-værdien afhænger af ikke blot den faktisk observerede stikprøve, men også af *hypotetiske* data under H_o. P-værdien angiver sandsynligheden for at få de observerede data *eller mere ekstreme data*, givet at H_o er sand. Man skal altså forestille sig, at eksperimentet gentages mange gange, og man ser da på, hvor ofte sådanne mere ekstreme observationer fremkommer under H_o. Hvis dette kun sker meget sjældent, er p-værdien lille og H_o forkastes.[2]

P-værdiens afhængighed af hypotetiske data under H_o bliver af ikke-frekvensbaserede statistikere opfattet som en afgørende svaghed. Et berømt citat er følgende (Jeffreys, 1961, p.385): "*What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.*" (kursiveringen er Jeffreys). Og Jeffreys fortsætter: "This seems a remarkable procedure."

Jeffreys' pointe kan illustreres med følgende eksempel (svarende til Example 30 i Berger and Wolpert, 1988, og Example 1 i Wagenmakers, 2007, der begge er inspireret af et tilsvarende eksempel fra Cox, 1958, p.368): Antag, at variablen X er fordelt som i nedenstående tabel, dvs. som enten $P_o(x)$ eller $P_1(x)$.

	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$	$x=6$
$P_o(x)$	0.04	0.30	0.31	0.31	0.03	0.01
$P_1(x)$	0.04	0.30	0.30	0.30	0.03	0.03

Lad $T(x) = x$ være teststatistikken for hypotesen at $P_o(x)$ er den sande fordeling, og antag at der observeres $x = 5$. Den klassiske p-værdi i dette test er $0.03 + 0.01 = 0.04$, dvs. hypotesen forkastes på 5% signifikansniveau. Det tilsvarende test for $P_1(x)$ giver en p-værdi på $0.03 + 0.03 = 0.06$, dvs. her forkastes hypotesen ikke.

[2] Fisher's (1925) oprindelige beskrivelse af p-værdien skete ikke indenfor rammerne af et eksplicit *repeated sampling* framework. Men i efterfølgende lærebogsfremstillinger defineres p-værdien (og lignende statistikker) typisk udfra tænkte gentagelser af det eksperimentelle setup. Fisher var i øvrigt en af foregangsmændene bag udviklingen af LP, uden dog at committe sig helt til princippet (Berger and Wolpert, 1988, p.22).

Bemærk, at $P(x = 5) = 0.03$ i begge tilfælde, dvs. den klassiske likelihood ratio (eller Bayes-faktor) er lig med 1 i begge tilfælde. Det *observerede* resultat ($x = 5$) er lige sandsynlig i de to fordelinger. Men p-værdien forkaster $P_o(x)$, mens $P_1(x)$ ikke forkastes, hvilket skyldes sandsynlighederne (0.01 og 0.03) for det *uobserverede* resultat $x = 6$. Med andre ord: $x = 6$ er ikke særlig sandsynlig under hverken $P_o(x)$ eller $P_1(x)$, og $x = 6$ blev da heller ikke observeret, men $P_o(x)$ forkastes mens $P_1(x)$ ikke forkastes, fordi $P_o(x)$ 'forudsiger' dette ikke-observerede resultat i *mindre* grad end $P_1(x)$.

Eksemplet illustrerer samtidig et andet særligt kendetegn ved p-værdien: p-værdien siger ingenting om den relative evidens for eller imod nul- vs. alternativhypotesen. I eksemplet er $x = 5$ lige (u)sandsynlig i begge fordelinger, så hvis de to fordelinger er de eneste mulige, giver $x = 5$ reelt ingen evidens imod $P_o(x)$ (Berger and Wolpert, 1988, p.109.3). I modsætning hertil giver Bayes-faktoren - som beskrives nedenfor - den relative evidens for H_o vs. H_1.

Findes der metoder, der overholder LP? Ja, der findes flere. En mindre skole indenfor statistik ('*likelihoodism*') anbefaler, at man vurderer evidensen i data alene ud fra likelihoodfunktionen uden brug af priors (som i den bayesianske tilgang) og uden brug af statistikker, der betinger på hypotetiske data (á la p-værdien).[3] Der findes også frekvensbaserede metoder, der stræber efter at overholde LP ('*conditional frequentism*'). Den største og mest indflydelsesrige klasse af metoder, der overholder LP, er dog den bayesianske. I det følgende vil jeg beskrive *Bayes-faktoren* og vise hvordan den overholder LP.

Lad $P(H_o)$ og $P(H_1) = 1 - P(H_o)$ være á priori sandsynlighederne for hhv. nul- og alternativhypotesen, og lad D være de observerede data (stikprøven). Bayes formel indebærer da følgende *posterior odds ratio* for de to hypoteser:

$$\frac{P(H_o \mid D)}{P(H_1 \mid D)} = \frac{P(D \mid H_o)}{P(D \mid H_1)} \cdot \frac{P(H_o)}{P(H_1)}. \tag{5}$$

$\frac{P(D|H_o)}{P(D|H_1)}$ er den relative likelihood, der i bayesiansk statistik kaldes for Bayes-faktoren (BF), og denne faktor angiver hvordan de observerede data ændrer á priori opfattelsen af hypoteserne til en ny (á posteriori) opfattelse. BF $>$ 1 ($<$ 1) indebærer, at vi i lyset af data opjusterer (nedjusterer) vores tro på H_o relativt til H_1.

[3] Maksimum likelihood (ML) *estimatoren*, eksempelvis, overholder LP. ML-estimatoren for andelsparameteren θ i binomialfordelingen er identisk med ML-estimatoren for θ i den negative binomialfordeling i afsnit 2.1 (for givet f).

Lad os vende tilbage til eksemplet i afsnit 2.1. Lad H_o: $\theta = \theta_o$, og H_1: $\theta \neq \theta_o$. Hvordan ser Bayes-faktoren ud i de to eksperimentelle setups? I binomialeksperimentet fås likelihoodværdien under H_o som $P(D \mid H_o) = P(D \mid \theta = \theta_o) = \binom{n}{s} \theta_o^s (1-\theta_o)^{n-s}$. For at beregne likelihoodværdien under alternativhypotesen, skal der integreres over de værdier θ kan antage under H_1: $P(D \mid H_1) = \int_0^1 P(D \mid \theta) f(\theta) d\theta$, hvor $f(\theta)$ er tæthedsfunktionen for θ. Et naturligt valg for $f(\theta)$ i dette tilfælde, hvor θ er en andel (sandsynlighedsparameter), er en betafordeling. Lad eksempelvis $\theta \sim$ Beta(1,1), dvs. en uniform fordeling i intervallet (0, 1). Bemærk at under H_1 er alle punkt-sandsynligheder i dette interval (inklusiv i punktet θ_o) lig med 0. Hermed fås $f(\theta) = 1$, og $P(D \mid H_1) = \binom{n}{s} \int_0^1 \theta^s (1-\theta)^{n-s} d\theta = \binom{n}{s} \cdot \frac{s! \cdot (n-s)!}{(n+1)!}$. Bayes-faktoren i binomialeksperimentet, BF_{bin}, kan dermed beregnes til:

$$\mathrm{BF}_{bin} = \frac{P(D \mid H_o)}{P(D \mid H_1)} = \frac{(n+1)! \cdot \theta_o^s (1-\theta_o)^{n-s}}{s!(n-s)!}. \tag{6}$$

I det negative-binomial setup fås $P(D \mid H_o) = \binom{n-1}{f-1} \theta_o^s (1-\theta_o)^{n-s}$ og $P(D \mid H_1) = \binom{n-1}{f-1} \int_0^1 \theta^s (1-\theta)^{n-s} d\theta = \binom{n-1}{f-1} \cdot \frac{s! \cdot (n-s)!}{(n+1)!}$. Hermed fås Bayes-faktoren, $\mathrm{BF}_{neg-bin}$, som:

$$\mathrm{BF}_{neg-bin} = \frac{(n+1)! \cdot \theta_o^s (1-\theta_o)^{n-s}}{s!(n-s)!}, \tag{7}$$

som er identisk med (6).

Bayes-faktoren er altså uafhængig af, om det præcise eksperimentelle setup er binomial aller negativ-binomial, så længe de underliggende bernoulli-forsøg er de samme, og med samme sandsynlighedsparameter θ, i de to eksperimenter. Dette gælder ikke blot i det konkrete eksempel her, men gælder generelt, og står i skarp kontrast til det frekvensbaserede approach, hvor - som vi så i afsnit 2.1 - p-værdien afhænger af, om det eksperimentelle setup er binomial eller negativ-binomial. Bayes-faktoren overholder dermed LP, hvilket grundlæggende skyldes, at BF kun afhænger af de *observerede* data gennem likelihoodfunktionen. Den klassiske p-værdi, derimod, afhænger af de observerede data *samt* hypotetiske data genereret under H_o i tænkte gentagelser af eksperimentet. Det er klart, at disse hypotetiske data afhænger af, om det eksperimentelle setup er binomial eller negativ-binomial.

Med den konkrete stikprøve i eksemplet i afsnit 2.1, $n = 12$ og $s = 9$, og med $\theta_o = 0.5$, fås $\mathrm{BF}_{bin} = \mathrm{BF}_{neg-bin} = \frac{13! \cdot 0.5^{12}}{9! \cdot 3!} = 0.70$. Hermed fører data altså til, at vi nedjusterer vores sandsynlighedsvurdering af H_o relativt til H_1. Hvis vi á priori havde en 'neutral' holdning til de to hypoteser, dvs. $P(H_o) = P(H_1) = 0.50$, fører dataanalysen til en *posterior odds ratio* på $\frac{P(H_o|D)}{P(H_1|D)} = \mathrm{BF} = 0.70$, og dermed til $P(H_o \mid D) = 0.41$ og $P(H_1 \mid D) = 0.59$.

Bemærk, at populationsparametre (her: θ) i bayesiansk analyse er stokastiske, i modsætning til den klassiske frekvensbaserede analyse, hvor populationsparametre opfattes som faste og ikke-stokastiske størrelser. I bayesiansk analyse antager man, at usikkerheden om populationsparametres værdier kan beskrives ved sandsynligheds(tætheds)funktioner. Det er dét der gør, at det giver mening at knytte sandsynligheder til hypoteserne H_o og H_1.

3.2 Tankevækkende implikationer af henholdsvis det frekvensbaserede approach og likelihoodprincippet.

Litteraturen er fuld af sjove eksempler på 'absurde' implikationer af hhv. LP og den frekvensbaserede tilgang, som de forskellige statistiske skoler slår hinanden oven i hovedet med. Kritikken mod den frekvensbaserede skole går især på afhængigheden af hypotetiske ikke-observerede data, jf. citatet fra Jeffreys (1961) i afsnit 3.1 vedr. p-værdien. Afhængigheden af hypotetiske data indebærer, at p-værdien afhænger af dataindsamlerens subjektive *intentioner*; intentioner, der i praksis ofte ikke er kendte for dem, der skal analysere de faktisk indsamlede data.

Problemet kan illustreres med et berømt eksempel fra Pratt (1962), der i forkortet form er følgende: En ingeniør har med et voltmeter indsamlet en stikprøve trukket fra en normalfordelt population. Alle observationer ligger mellem 75 og 99. Ingeniøren ønsker at foretage inferens om middelværdien og beder en frekvensbaseret statistiker om hjælp. Statistikeren hæfter sig ved observationen 99 og spørger, om voltmeteret mon er censureret. Ingeniøren bekræfter, at måleinstrumentet kun går til 100, men da alle observationer jo var under 100, burde det ikke være noget problem, og hvis han havde fået en måling på 100, ville han have skiftet til et instrument, der kan måle til 1000. Statistikeren stiller sig tilfreds med denne forklaring og beregner et klassisk 95% konfidensinterval for middelværdien. Næste dag ringer ingeniøren til statistikeren og fortæller, at han lige har opdaget, at instrumentet, der går til 1000, er i stykker og at han derfor ikke ville kunne have brugt det. Til ingeniørens store overraskelse modtager statistikeren denne oplysning med besked om, at det faktisk ændrer det eksperimentelle setup, og at den statis-

tiske analyse nu afhænger af, om ingeniøren - havde han vidst at instrumentet var i stykker - ville have ventet med at foretage målingerne på instrumentet der går til 100 indtil instrumentet der går til 1000 var blevet repareret. Hvis ikke, skal analysen tage højde for, at fordelingen for observationerne nu de facto er trunkeret. Ingeniøren forstår det ikke og udbryder: "Men resultatet af eksperimentet blev jo det samme som hvis instrumentet ikke var i stykker, så hvordan kan det påvirke den statistiske analyse?".

Pratt erklærer sig enig med ingeniøren: "*I agree with the engineer. If the sample has voltages under 100, it doesn't matter whether the upper limit is 100, 1000, or 1 million. The sample provides the same information in any case. And this is true whether the end-product of the analysis is an evidential interpretation, a working conclusion, or an action.*" (Pratt, 1962, p.315). Pratt advokerer følgelig for, at man bør anvende metoder, der overholder LP, hvorved man undgår sådanne paradokser (se Berger and Wolpert, 1988, Example 24, for yderligere diskussion af dette eksempel).

En problemstilling, der i særlig grad har optaget sindene, er det såkaldte 'stop-regel princip' (*Stopping Rule Principle*, SRP), der siger, at *årsagen* til at stoppe dataindsamlingen i et eksperiment er uden betydning for den statistiske inferens om en given parameter, så længe denne årsag ikke påvirker likelihoodfunktionen. SRP kaldes også for *optional stopping*, og følger naturligt at LP. Vi så i afsnit 3.1, at Bayes-faktoren er uafhængig af, om det eksperimentelle setup er binomial eller negativ-binomial, hvilket indebærer, at inferens baseret på Bayes-faktoren er uafhængig af, om stop-reglen er at sample indtil man har fået n observationer, eller indtil man har fået f fiaskoer. Mere generelt indebærer SRP, at "*Experiments may end because the data looks convincing enough, because money runs out, or because the experimenter has a dinner date.*" (Berger and Wolpert, 1988, p.77).

Stop-regel princippet er naturligvis helt uantagelig for frekventister, der ikke er blege for at kalde det for svindel. Af og til afvises LP alene med henvisning til stop-regel princippet. Og det er klart, at princippet er helt uforenelig med frekventistiske principper. Indenfor rammerne af det frekvensbaserede approach, er *optional stopping* ren 'p-hacking'. Det kan vises, at for et positivt men arbitrært lille signifikansniveau, vil den klassiske p-værdi altid føre til 'signifikans' selvom H_o er sand, hvis man blot bliver ved med at indsamle data og for hver nyt datapunkt tester hypotesen (Berger and Wolpert, 1988, p.77; Wagenmakers, 2007, p.785). Dette skyldes grundlæggende, at p-værdien er uniformt fordelt, $U(0, 1)$, under H_o.

Årsagen til at stop-regel princippet gælder under LP, er, at stop-regler

ikke påvirker likelihoodfunktionen. Indenfor rammerne af det bayesianske approach, er *optional stopping* tilladeligt. Eksempelvis en stop-regel, der siger, at man sampler indtil Bayes-faktoren for Model-1 vs. Model-2 når en præspecificeret værdi. I modsætning til det frekvensbaserede approach baseret på p-værdien, fører en sådan strategi ikke automatisk til valg af en given model, og hvis H_o er sand, går Bayes-faktoren for H_o vs. H_1 mod uendelig når stikprøvestørrelsen vokser.

3.3 Det matematiske grundlag for likelihoodprincippet.

Birnbaum (1962) etablerede det matematiske grundlag for LP. Han viser, at LP følger af to andre veletablerede principper i statistik: 'Sufficiensprincippet', SP (*principle of sufficiency*) og 'conditionality-princippet' (*principle of conditionality*), CP. Kort fortalt siger SP, at evidensen fra et eksperiment fuldt kan opsummeres ved sufficiente statistikker, mens CP siger, at kun det udførte eksperiment er relevant for evidensen.

Mere formelt beskriver Birnbaum disse principper som følger: Lad E være den matematisk-statistiske model for et givet eksperiment, og $S = \{x\}$ det tilhørende mulige udfaldsrum for den stokastiske variabel X, hvis statistiske fordeling er beskrevet ved parameteren θ. Definer endvidere Ev(E, x) til at være 'evidensen' (*evidential meaning*), der kan udledes fra eksperimentet. Lad $t(x)$ være en vilkårlig sufficient statistik, og lad E' være et afledt eksperiment, der har samme parameterrum som E, sådan at hvis et vilkårligt outcome x i E observeres, da observeres samtidig det afledte outcome $t = t(x)$ i E'. I så fald gælder for ethvert x, at Ev(E, x) = Ev(E', t), hvor $t = t(x)$. Dette kaldes for 'sufficiensprincippet', SP. En implikation af SP er, at hvis to outcomes, x og x', fra et vilkårligt eksperiment E har den samme likelihoodfunktion (rettere: de to likelihoodfunktioner er proportionale), da gælder Ev(E, x) = Ev(E, x').

Principle of conditionality, CP, beskriver Birnbaum på følgende måde: Lad E være et 'mixtur-eksperiment' med komponenter $\{E_h\}$, der består af to steps: I første step trækkes en observation fra en stokastisk variabel, H, med en kendt fordeling, der ikke afhænger af θ. I næste step udføres eksperimentet E_h, hvilket giver outcome x_h; dvs. det første step bestemmer hvilket konkret eksperiment der skal udføres i det andet step (tænkt på det som at man slår plat eller krone om hvilket af to mulige eksperimenter, der skal udføres, eksempelvis binomial- eller negativ-binomial eksperimentet ovenfor). Hvis et eksperiment E består af et mixtur-eksperiment med komponenter $\{E_h\}$ med mulige outcomes (E_h, x_h), da gælder Ev(E, (E_h, x_h)) = Ev(E_h, x_h).

Birnbaum beskriver 'likelihoodprincippet', LP, som følger: Lad $f(x, \theta)$ og $g(y, \theta)$ være likelihoodfunktionerne for outcomes x og y i henholdsvis eksperiment E og eksperiment E', og hvor $f(x, \theta) = cg(y, \theta)$, hvor $c = c(x, y)$ er en vilkårlig positiv konstant. Da gælder $\mathrm{Ev}(E, x) = \mathrm{Ev}(E', y)$. Birnbaum viser matematisk, at LP implicerer - og impliceres af - SP og CP (Birnbaum, 1962, Lemma 2). Birnbaum (1962, p.271) beskriver uformelt CP som "*the irrelevance of (component) experiments not actually performed*" og LP som "*the irrelevance of outcomes not actually observed*".

Sufficiensprincippet (SP) og conditionality-princippet (CP) virker måske hver især plausible, og dog afviser mange statistikere likelihoodprincippet (LP) med henvisning til dets 'absurde' implikationer (jf. afsnit 3.2). LP er kontroversiel, og der har i tidens løb været gjort flere forsøg på at tilbagevise Birnbaum's bevis for at SP og CP indebærer LP (se Berger and Wolpert, 1988, for detaljer), men der er i litteraturen ikke enighed om validiteten af disse forsøg. Mayo (2014) med efterfølgende diskussion, samt Pena and Berger (2017), er nyere bidrag til diskussionen.

Berger and Wolpert (1988) fremfører gentagne gange (Example 2 og pp. 25-26, 45 og 65-66), at CP i sig selv strider mod det frekvensbaserede princip om inddragelse af hypotetiske ikke-observerede data. I relation til mixtur-eksperimentet ovenfor, følger man de frekvensbaserede principper strengt, bør den statistiske inferens inddrage de mulige udfald i første step (se også Cox, 1958, pp.360-361, for diskussion af mixtur-eksperimentet).

Det er bemærkelsesværdigt, at der i dag i den statistiske litteratur så mange år efter Birnbaum (1962) stadig ikke er konsensus om det matematiske grundlag for LP. Ergo må man forlade sig på sin intuition i bedømmelsen af LP og dets praktiske implikationer.

3.4 Implikationer for samfundsvidenskabelig praksis.

De fleste empiriske samfundsforskere vil i svage øjeblikke erkende, at de har bedrevet 'p-hacking', dvs. torturcret (testet) data til de bekender (giver signifikans), og dermed givetvis medvirket til de mange *false discoveries* (jf. indledningen til denne artikel). Med mindre man direkte har manipuleret eller svindlet med data, er en sådan procedure ikke nødvendigvis udtryk for svindel eller videnskabelig uredelighed. Men det er under alle omstændigheder ikke i overensstemmelse med de grundlæggende principper bag de klassiske frekvensbaserede metoder. Stop-regel princippet (*optional stopping*), som beskrevet i afsnit 3.2, er totalt no go i klassisk statistik, men ikke desto mindre bedriver vi optional stopping - og lignende data mining - hele tiden!

Berger and Wolpert (1988, p.74) argumenterer for, at "*The theoretical and practical implications of the [Stopping Rule Principle] to such fields as sequential analysis and clinical trials are enormous.*" I den slags forsøg er det ikke ualmindeligt at analysere data efterhånden som de indløber. Dette behøver ikke være et problem, hvis de metoder man anvender overholder likelihoodprincippet.

Et konkret eksempel indenfor samfundsvidenskab er opinionsundersøgelser, eksempelvis de politiske meningsmålinger, som analyseinstitutter jævnligt udsender baseret på stikprøver på ca. 1000 tilfældigt udvalgte personer. I sådanne opinionsundersøgelser rapporteres statistiske usikkerheder typisk beregnet udfra klassiske 95% konfidensintervaller, hvor man antager en fast og på forhånd kendt stikprøvestørrelse (á la binomialeksperimentet i afsnit 2.1), på trods af, at den eksakte stikprøvestørrelse i disse undersøgelser som regel ikke er kendt på forhånd. Et yderligere særligt kendetegn ved klassiske konfidensintervaller er, at de ikke indeholder fordelingsmæssig information om værdierne indenfor intervallet, et forhold som klassiske frekventister naturligvis altid har været klar over kan være en begrænsning: "*the method of confidence intervals, as usually formulated, gives only one interval at some preselected level of probability. ... For when we write down the confidence interval ... for a completely unknown normal mean, there is certainly a sense in which the unknown mean θ is likely to lie near the centre of the interval, and rather unlikely to lie near the ends and in which, in this case, even if θ does lie outside the interval, it is probably not far outside. The usual theory of confidence intervals gives no direct expression of these facts.*" (Cox, 1958, p.363).

Som et alternativ (eller supplement) til det klassiske konfidensinterval, kan den statistiske usikkerhed beregnes ud fra bayesianske '*posterior intervals*' (også kaldet for '*credible intervals*' eller '*highest density intervals*') baseret på á posteriori fordelingen for populationsparemeteren, hvilket i en meningsmåling vil sige andelen, der vil stemme på det givne politiske parti. En oplagt prior fordeling for andelen vil være en passende betafordeling centreret omkring det seneste valgresultat for partiet, og med en spredning, der afspejler variationen i partiets tilslutning i de seneste målinger. Udover at overholde LP, er der flere fordele ved et sådant *posterior interval*. For det første forudsætter det ikke en kendt og forudbestemt stikprøvestørrelse. For det andet rapporterer det direkte - for den givne stikprøve - det interval, populationsparameteren (her: stemmeandelen) med x% sandsynlighed ligger indenfor (i modsætning til det klassiske konfidensinterval, der skal fortolkes

indenfor rammerne af et frekvensbaseret *repeated sampling* approach, jf. afsnit 2). Og for det tredje giver *posterior intervals* sandsynlighedsmæssig information om andelene indenfor intervallet. Andele tæt på modus i posterior fordelingen er mere sandsynlige end andele langt fra modus. Dette er i modsætning til det klassiske konfidensinterval, hvor det - jf. citatet fra Cox (1958) ovenfor - ikke giver mening at sige noget om sandsynligheden for enkelte værdier indenfor intervallet (se Kruschke and Liddell, 2018, for yderligere diskussion af denne problemstilling).

Uanset hvilken metodologi man anvender, forudsætter valid statistisk inferens, at den statistiske model giver en adækvat beskrivelse af stikprøven. I kontrollerede eksperimenter har man i sagens natur bedre mulighed for at kontrollere stikprøvens egenskaber, eksempelvis sikre uafhængighed mellem observationerne. Men som tidligere diskuteret, er samfundsvidenskabelige data ofte ikke frembragt ved kontrollerede eksperimenter. Faktisk vil det typisk være sådan, at det netop er den data-genererende process man jagter. Eksempelvis empiriske studier af prisdannelsen på de finansielle markeder. I sådanne studier er der ikke et kontrolleret eksperimentelt setup, men derimod et passivt observeret datamateriale indeholdende finansielle priser og observationer på diverse relevante baggrundsfaktorer, der kan tænkes at 'forklare' priserne. Her kender vi ikke den underliggende data-genererende process; i stedet er det den data-genererende process bag prisdannelsen vi ønsker at afdække.

Frekventister vil her typisk anvende klassiske fejlspecifikationstests, baseret på modellens residualer, til at tjekke adækvansen. Statistikere, der henholder sig til LP, afviser en sådan tilgang, da residualer ikke er baseret på sufficiente statistikker, hvorved sådanne fejlspecifikationstests ikke overholder LP. Omvendt er statistikere, der bekender sig til LP - eksempelvis bayesianere -, blevet kritiseret for ikke at have fokus på at tjekke modelantagelserne (se Mayo, 2014, p.230).

Videnskabsfilosofisk tilhører den frekvensbaserede tilgang det Popperske '*hypothetico-deductive*' princip, efter hvilket videnskabelige teorier - hypoteser - forsøges falsificeret gennem konfrontation med data. Klassiske fejlspecifikationstests på en models residualer kan siges at være et sådant falsifikansredskab. En model, der overlever batteriet af fejlspecifikationstests, accepteres som 'sand', mens en model, der ikke overlever disse tests, afvises som 'falsk'. I modsætning hertil står den bayesianske tilgang, hvor modeller - hypoteser - ikke opfattes som 'sande' eller 'falske'. Modeller tildeles á priori sandsynligheder, der gennem data (likelihoodfunktionen) modificeres til á posteriori

sandsynligheder. Modeltjek spiller med andre ord fundamentalt forskellige roller i de to tilgange. Men uanset hvad, er det i begge tilgange naturligvis vigtigt at sikre, at den antagede statistiske fordeling for de observerede data (likelihoodfunktionen) er en rimelig approksimation. Dette er selvsagt lettere at sikre, når data er frembragt i kontrollerede eksperimenter, end når data er ikke-eksperimentelle og passivt observerede, som det ofte er tilfældet indenfor samfundsvidenskab, og hvor den underliggende økonomiske (eller politologiske, sociologiske, psykologiske, etc.) model under alle omstændigheder jo blot er en grov approksimation til virkeligheden.

Klassiske statistikere har også kritiseret den bayesianske tilgang for ikke at have fokus på at kunne kontrollere sandsynlighederne for Type-1 og Type-2 fejl. I den klassiske Neyman-Pearson tilgang udtrykker disse fejlsandsynligheder langsigts-frekvenser ved *repeated sampling* og fortæller i sig selv ikke noget om den givne stikprøve. Sådanne fejlsandsynligheder spiller i sagens natur ikke den samme rolle i bayesiansk analyse, hvor fokus netop er på den givne stikprøve, og hvor *repeated sampling* princippet ingen rolle spiller.

Berger and Wolpert (1988, p.73) påpeger endvidere, at kontrol over de klassiske Type-1 og Type-2 fejlsandsynligheder ikke indebærer kontrol over fejlsandsynlighederne når man forkaster henholdsvis ikke-forkaster nulhypotesen. Eksempelvis for et test med lav styrke (høj sandsynlighed for Type-2 fejl), vil sandsynligheden for at H_o er sand, hvis man forkaster H_o, være langt højere end sandsynligheden for at forkaste en sand H_o, dvs. Type-1 fejlsandsynligheden (signifikansniveauet). Og i de klassiske tests kan man ikke kontrollere både Type-1 fejlen og Type-2 fejlen samtidig. Sædvanligvis kontrollerer man Type-1 fejlen, og håber så på, at Type-2 fejlsandsynligheden ikke er alt for høj. Dvs. man har ingen kontrol over, hvor ofte man begår en fejl, når man forkaster hypotesen (Sellke et al., 2001, indeholder yderligere diskussion og illustration af denne problemstilling). Harvey (2017) argumenterer for, at fundet af mere end 300 risikofaktorer for aktiemarkedet i den empiriske finansieringslitteratur over de seneste årtier delvist er et resultat af dette problem. Hovedparten af disse fundne risikofaktorer er givetvis *false discoveries*.

4. Konklusion.

I den ideelle verden kan vi designe, fuldt kontrollere og gentage det eksperimentelle setup, der ligger til grund for udtrækning af stikprøver. Men vi

lever ikke i en ideel verden. Ofte er de data vi arbejder med ikke eksperimentelle, og tanken om at gentage 'eksperimentet' er ofte svært at forestille sig. Selv med eksperimentelle data har vi ofte ikke fuld kontrol over - eller kendskab til - det præcise eksperimentelle setup. Da de klassiske frekvensbaserede statistiske metoder, som de fleste af os er oplært i, anvender i vores forskning og selv underviser i, bygger på et fuldt specificeret eksperimentelt setup og '*the repeated sampling framework*', hvor statistikkers langsigtede frekvensmæssige egenskaber i tænkte gentagelser af eksperimentet spiller en vigtig rolle, giver anvendelsen af disse metoder anledning til metodologiske og videnskabsteoretiske overvejelser. Er de statistiske metoder vi anvender, grundlæggende velegnede til analyse af den type data vi arbejder med?

I denne artikel har jeg argumenteret for, at metoder der - i modsætning til de frekvensbaserede metoder - bygger på 'likelihoodprincippet' (LP), måske er bedre egnede til de data vi som samfundsforskere arbejder med. Metoder der overholder LP, bygger ikke på et *repeated sampling* framework med tænkte gentagelser af eksperimentet, men alene på den givne stikprøve (likelihoodfunktionen). Dette giver i sammenligning med den klassiske tilgang visse frihedsgrader i forhold til det eksperimentelle setup, især ved analyse af passivt observerede ikke-eksperimentelle data. Under LP, vil forhold, der ikke påvirker likelihoodfunktionen (stop-regler, fast vs. ikke-fast stikprøvestørrelse, hvor mange tests der udføres, etc.), ikke have betydning for inferensen.

Anvendelse af metoder, der overholder LP, forudsætter naturligvis, at der kan opstilles en likelihoodfunktion og at den antagede fordeling er en god approksimation til dataenes fordeling. I tilfælde, hvor dette ikke lader sig gøre, og man eksempelvis må benytte sig af ikke-parametriske metoder, kan LP i sagens natur ikke fungere som princip for den statistiske analyse. Men vælger man den klassiske frekvensbaserede tilgang, er det vigtigt at være opmærksom på de ret strenge antagelser vedrørende det eksperimentelle setup, der kendetegner denne tilgang. Det er mit indtryk, at vi samfundsforskere har det med at glemme - eller se stort på - disse antagelser.

Referencer.

Berger, J.O. and R.L. Wolpert (1988): *The Likelihood Principle.* Lecture Notes - Monograph Series (Volume 6), Institute of Mathematical Statistics, Hayward, California.

Birnbaum, A. (1962): On the foundations of statistical inference. *Journal of the Americal Statistical Association* 57, 269-306.

Cox, D.R. (1958): Some problems connected with statistical inference. *Annals of Mathematical Statistics* 29, 357-372.

Cox, D.R. and D.V. Hinkley (1974): *Theoretical Statistics*. Chapmann and Hall, London.

Engsted, T. (2019): Bayesianske hypotesetests. I: *Symposium i Anvendt Statistik 2019* (red. Peter Linde), Københavns Universitet og Det Nationale Forskningscenter for Arbejdsmiljø.

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, London.

Kruschke, J.K. and T.M. Liddell (2018): The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review* 25, 178-206.

Leamer, E.E. (1978): *Specification Searches: Ad Hoc Inference With Nonexperimental Data*. John Wiley & Sons, New York.

McShane, B.B., D.G. Gal, A. Gelman, C. Robert, and J.L. Tackett (2019): Abandon statistical significance. *American Statistician* 73, 235-245.

Mayo, D.G. (2014): On the Birnbaum argument for the strong likelihood principle (with discussion). *Statistical Science* 29, 227-266.

Neyman, J. and E.S. Pearson (1933): On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society Series A 231*, 289-337.

Pena, V. and J.O. Berger (2017): A note on recent criticisms to Birnbaum's theorem. *Working Paper* (https://arxiv.org/abs/1711.08093).

Pratt, J.W. (1962): Discussion of 'On the foundations of statistical inference'. *Journal of the Americal Statistical Association* 57, 314-316.

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

Wagenmakers, E.-J. (2007): A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review* 14, 779-804.

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

Wasserstein, L.R., A.L. Schirm, and N. Lazar (2019): Editorial: Moving to a world beyond "p<0.05". *American Statistician* 73, 1-19.

Inferens i mixed models i R - hinsides det sædvanlige likelihood ratio test

Søren Højsgaard*

6. januar, 2020

Resumé: _Inferens i lineære mixed models og i generaliserede lineære mixed models er ofte baseret på χ^2 approximationen til likelihood ratio teststørrelsens fordeling. Det går som regel godt i store datasæt, men et datasæt kan på samme tid være stort med hensyn til nogle aspekter af en problemstilling og lille med hensyn til andre aspekter. Et klassisk eksempel er data fra et split-plot forsøg: Delploteffekten kan være velbestemt mens helploteffekten ofte vil være dårligere bestemt. I visse planlagte forsøgstyper ved vi, hvordan vi skal håndtere hypotesetests i sådanne modeller. I observationelle studier er det mindre klart, hvordan man skal håndtere hypotesetests. Een mulighed mulighed er at lave en form for F-test hvor nævner-frihedsgraderne er justerede (typisk) for at tage hånd om, at dispersionsparametre er estimerede fra data og dermed ikke må betragtes som kendte. En anden tilgang er at basere tests på parametrisk bootstrap. Fordelen ved denne metode er, at den umiddelbart lader sig anvende i mere generelle situationer end lineære mixed models; f.eks. i generaliserede lineære modeller. Begge metoder er tilgængelige i R pakken `pbkrtest`.

1 Introduktion

Mixed models håndteres i `R` (R Core Team 2019), oftest med `lme4` pakken (Bates et al. 2015). Tests er baserede på χ^2 approksimationen af likelihood ratio (LR) teststørrelsen, hvilket fungerer fint i store datasæt men ofte mindre godt i små datasæt. Dertil kommer, at et datasæt kan være stort med hensyn til nogle aspekter af en model og samtidig lille med hensyn til andre aspekter. R-pakken `pbkrtest` tilbyder alternativer til χ^2 approksimationen af LR teststørrelsen, nemlig: 1) Tests baserede på en F-teststørrelse (hvor nævnerfrihedsgraderne estimeres fra data), 2) tests baserede på parametrisk bootstrap (hvor data er simuleret under modellen). Parametrisk bootstrap kan også bruges for tests i generaliserede lineære modeller. Med (lineære) mixed models forstås i det følgende modeller af formen

$$y = X\beta + Zu + e$$

hvor y og e er n vektorer af stokastiske variable, X er $n \times p$ model matrix for systematiske effekter, β er p vektor af regressionskoefficienter, Z er $n \times q$ model matrix for de tilfældig effekter og u er q vector af tilfældige effekter. Det antages at $u \sim N(0, G)$ og $e \sim N(0, R)$ og at u og e er uafhængige. Generaliserede lineære modeller modeller er som generaliserede lineære modeller at det antages at $g(\mu) = X\beta + Zu$, hvor g er linkfunktionen.

*University of Aalborg, Denmark

Tabel 1: Simuleret datasæt. y_1 er en numerisk respons.

	y1	y2	grp	subj	y1	y2	grp	subj
2	67	1	ctrl	subj1	26	2	trt1	subj4
3	72	1	ctrl	subj1	45	1	trt1	subj4
1	140	1	ctrl	subj1	90	0	trt1	subj4
4	13	1	ctrl	subj2	48	2	trt1	subj5
6	27	2	ctrl	subj2	53	3	trt1	subj5
5	37	1	ctrl	subj2	95	2	trt1	subj5
8	-76	0	ctrl	subj3	70	2	trt1	subj6
7	-66	2	ctrl	subj3	99	0	trt1	subj6
9	-56	3	ctrl	subj3	131	0	trt1	subj6

2 Eksempel: Dobbeltregistreing i laboratorieforsøg

Betragt et konstrueret, men meget simpelt, eksempel: Vi ønsker at sammenligne to grupper (f.eks. behandling mod kontrol). Til rådighed er der M *units* (petriskåle, personer, dyr...) per gruppe og der måles på hver unit ialt R gange. Målinger på samme unit vil oftest give anledning til *clustering* i data. Et datasæt i "long format" vil altså have $T = 2 \times M \times R$ rækker. Table 1 viser et simuleret datasæt med $M = 3$ units per gruppe; $R = 3$ gentagne målinger per unit. Problemstillingen er, at målinger på samme unit typisk er positivt korrelerede, og hvis man ikke tager højde for dette, så kan man komme til at overvurdere mængden af information, der er i datasættet. Mere konkret er det typiske billede at 1) estimater for standardfejl blive for små og 2) derfor bliver teststørrelser bliver for store og 3) derfor bliver p-værdier for små og 4) derfor kommer effekter til at fremstå stærkere end de i virkeligheden er.

Ignorer clustering i data: En simpel regressionsmodel er

$$y_{gir} = \mu + \beta_g + e_{gir},$$

hvor g refererer til gruppe, i til individ indenfor gruppe, r til måling indenfor individ og $e_{gir} \sim N(0, \sigma^2)$. Denne model vil typisk være utilstrækkelig fordi man har målt på samme unit flere gange, men vi inkluderer resultatet for sammenligningens skyld. Behandlingseffekten er $\beta_{trt} - \beta_{ctrl}$ og denne omtales i det følgende som β_1 i benævnes i tabeller med `grptrt1`. Estimatet for β_1 giver dermed behandlingseffekten. Tabel 2 viser resultatet af at fitte denne model. p-værdien for behandlingseffekten bliver meget lille, hvilket indikerer stor sikkerhed af en behandlingseffekt.

Tabel 2: Resultatet af at analysere data når clustering i units ignoreres: p-værdien for behandlingseffekten er ganske lille, hvilket indikerer en behandlingseffekt.

	Estimate	Std. Error	t value	Pr(>|t|)	Pr(>X2)
(Intercept)	17.6	18.8	0.934	0.364	0.350
grptrt1	55.4	26.6	2.086	0.053	0.037

Standard tilgang: Analyser gennemsnit: En hyppigt anvendt tilgang er at udregne gennemsnit for hver unit og analysere disse. Mere konkret betragtes modellen

$$\bar{y}_{gi} = \mu + \beta_g + e_{gi},$$

hvor der er beregnet gennemsnit indenfor hver unit. Denne tilgang virker fint i den forstand, at man får de rette tests når data er balancerede. Man får dog ikke noget estimat for "within-subject" variationen, hvilket dog ikke nødvendigvis er et stort problem i den konkrete sammenhæng. At analysere gennemsnitte er dog langt fra altid en mulighed. Tabel 3 indeholder resultatet for analyse af gennemsnittet. Tabellen indeholder både resultaterne for tests baseret på t-fordelingen (svarende til at der tages højde for, at variansen er estimeret fra data) og for tests baseret på normalfordelingen (svarende til at der ikke tages højde for, at variansen er estimeret fra data).

Tabel 3: Resultat efter analyse af gennemsnit over units.

	Estimate	Std. Error	t value	Pr(>|t|)	Pr(>X2)
(Intercept)	17.6	34.0	0.516	0.633	0.606
grptrt1	55.4	48.1	1.152	0.314	0.249

Model med tilfældige effekter: At analysere et gennemsnit er muligt i dette eksempel men langt fra altid. Et alternativ er at anvende en lineær mixed model (i dette tilfælde en varianskomponent model) hvor unit optræder som en tilfældig effekt. Dvs. vi betragter modellen

$$y_{gir} = \mu + \beta_g + U_{gi} + e_{gir},$$

hvor $U_{gi} \sim N(0, \omega^2)$ og $e_{gir} \sim N(0, \sigma^2)$. Resultatet ses i Tabel 4. Bemærk at testet er baseret på χ^2 fordelingen. Det vil sige at der tages ikke højde for, at variansen er estimeret. I stedet er der en implicit antagelse om, at den estimerede varians er lig med den sande varians. Bemærk at p-værdierne er de samme som p-værdierne baseret på χ^2 approksimationen i Table 3. Dette er en konsekvens af, at data er balancerede.

Tabel 4: Resultat efter at fitte en mixed model med unit som tilfældig effekt.

term	estimate	std.error	statistic	Pr(>X2)
(Intercept)	17.6	34.0	0.516	0.606
grptrt1	55.4	48.1	1.152	0.249

3 Muligheder med `pbkrtest` pakken:

R pakken `pbkrtest` implementerer to methoder for modelsammenligninger i mixed models, hvori der tages højde for at varians- og kovariansparametre er estimerede fra data: 1) Parametrisk boostrap og 2) Kenward-Rogers approksimation (deraf navnet på pakken).

3.1 Kenward & Rogers tilgang

I kort form er tilgangen i Kenward and Roger (1997) som følger: For den multivariate normalfordeling $Y \sim N(X\beta, \Sigma)$ betragtes test af hypotesen $L(\beta - \beta_0) = 0$. Da $\hat{\beta} \sim N_d(\beta, \Phi)$ bliver en Wald test-størrelse

$$W = [L(\hat{\beta} - \beta_0)]'[L\Phi L']^{-1}[L(\hat{\beta} - \beta_0)].$$

Asympototisk er $W \sim \chi^2_d$-fordelt under null hypotesen. For at beregne denne størrelse skal et estimat $\hat{\Phi}$ anvendes. Implicit i antagelsen om at W skal være asymptotisk χ^2_d fordelt er at $\hat{\Phi}$ er lig med den sande varians. En skaleret version af W er

$$F = \frac{1}{d}W = \frac{1}{d}(\hat{\beta} - \beta_0)'L'(L'\Phi(\hat{\sigma})L)^{-1}L(\hat{\beta} - \beta_0).$$

I beregningen af F er $\Phi(\sigma) = (X'\Sigma(\sigma)X)^{-1} \approx \mathbb{C}ov(\hat{\beta})$, $\hat{\sigma}$ er vektor af REML estimater for elementerne i $\Sigma = \mathbb{V}ar(Y)$ og $\hat{\beta}$ er REML estimate for β.

Asymptotisk er $F \sim \frac{1}{d}\chi^2_d$ under null hypotesen, og man man kan tænke på F som grænsen af en $F_{d,m}$-fordeling når $m \to \infty$ Een måde hvorpå man kan tage højde for at $\Phi = \mathbb{V}ar(\hat{\beta})$ er estimeret fra data er ved at komme med et bedre bud på hvad nævnerfrihedsgraderne m er (bedre bud end $m = \infty$). Kenward and Roger (1997) gjorde følgende:

- Erstattede Φ med en forbedret small-sample approksimation Φ_A.
- Udledte formler for middelværdi E^* and varians V^* af F (baseret på en førsteordens Taylorudvikling).
- Skalerede F med en faktor λ og bestemte nævner frihedsgraderne m ved at match momenterne af F/λ med momenterne i en $F_{d,m}$ fordeling.

Anvendelse af Kenward-Rogers metode: Tabel 5 viser resultat efter at fitte en mixed model med unit som tilfældig effekt til de simulerede data. Den rapporterede p-værdi er for testet af ingen effekt at behandling. Testet er baseret på at approksimere teststørrelsen med en F-størrelse, hvori frihedsgraderne er estimerede udfra data. Bemærk, at p-værdien er den samme p-værdien i Tabel 3, hvor det er gennemsnittene, der analyseres.

Tabel 5: Resultat efter at fitte en mixed model med unit som tilfældig effekt. Den rapporterede p-værdi er for testet af ingen effekt at behandling. Testet er baseret på at approksimere teststørrelsen med en F-størrelse, hvori frihedsgraderne er estimerede udfra data.

	statistic	ndf	ddf	F.scaling	p.value
Ftest	1.33	1	4	1	0.314

3.2 Parametrisk bootstrap

Tilgangen i parametrisk bootstrap er som følger. Vi betragter to konkurrerende modeller: En stor model $f_1(y;\theta)$ og en simplere null model $f_0(y;\theta_0)$; null-modellen er en delmodel af den store model. Vi beregner en

teststørrelse t_{obs}. Så bliver p-værdien for hypotesen

$$p = \sup_{\theta \in \Theta_0} Pr_\theta(T \geq t_{obs}),$$

hvor supremum er under hypotesen. Sædvanligvis kan man ikke beregne dette supremum i praksis, så i stedet beregner vi teststandsynligheden baseret på parameterestimatet, dvs.

$$p^{PB} = Pr_{\hat{\theta}}(T \geq t_{obs}),$$

I praksis approksimeres p^{PB} som følger:

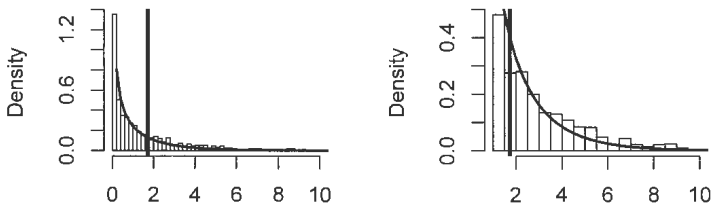
1. Træk B parametrisk bootstrap datasæt $D^1, \ldots D^B$ fra den fittede null model $f_0(\cdot; \hat{\theta}_0)$.
2. Fit den store og null modellen til hvert af disse datasæts.
3. Beregn likelihood ratio (LR) teststørrelsen for hvert simuleret datasæt. Dette giver referencefordelingen.
4. Beregn hvor ekstrem den observerede teststørrelse er; dette giver p-værdien.

Resultatet er anvendelsen af metoden er vist i Table 6.

Tabel 6: Resultat efter at fitte en mixed model med unit som tilfældig effekt. Den rapporterede p-værdi er for testet af ingen effekt at behandling og er beregnet ved parametrisk bootstrap.

	statistic	ndf	ddf	F.scaling	p.value
Ftest	1.33	1	4	1	0.314

Figur 1 viser χ^2_1 fordelingen (kurve) lagt oven på simuleret reference fordeling. Den simulerede referencefordeling har tungere hale end den teoretiske, og dette giver den større p-værdi.



Figur 1: Tætheden for den approximerende χ^2 fordeling lagt ovenpå den referencefordeling man får ved parametrisk bootstrap, dvs. et histogram. Til venstre: Hele intervallet for teststørrelsen. Til højre: Den del af halen af fordelingen det er relevant at betragte.

Parametrisk bootstrap er en computerintensiv metode, men der er en række muligheder for at gøre beregningerne hurtigere:

1. Seventielle p-værdier: Ovenfor simulerede man et fast antal værdier $t^1, \ldots, t^B$ af teststørrelsen under hypotesen for at kunne beregne p^{PB}. Et alternativ er at kan man i stedet introducere en stop-regel,

f.eks. *Simuler indtil vi har opnået f.eks. $h = 20$ værdier af t^j, der er større end t_{obs}.* Hvis dette er opnået efter J simulationer, så skal den rapporterede p-værdi være h/J.

2. Parallelle beregninger: En anden måde at gøre beregningerne hurtigere på er ved at udnytte flere kerner på samme computer. Dette sker som default på linux og mac platforme; på windows platform skal gå igennem visse opsætningsskridt.

3. Parametrisk form af referencefordelingen: Estimation af hale-sandsynligheder kræver flere samples en at estimere middelværdi og varians af fordelingen. Derfor er det fristende at approximere en simuleret referencefordeling med en kendt fordeling så færre simulationer er nødvendige. Eksempelvis kan man matche middelværdi og varians i en gammafordeling med middelværdi og varians af den simulerede referencefordeling og derefter beregne halvsandsynligheder i denne gammafordeling.

4 Simulationsstudium

Betragt igen situationen i Afsnit 2. Vi ønsker at teste hypotesen at der ikke er nogen behandlingseffekt (forskel i middelværdi). Vi gentager studiet mange gange (f.eks. 1000 gange). Da studierne er lavet ved computersimulation kan vi generere data således, at vi ved at der ikke er nogen behandlingseffekt. Hvis der ikke er nogen behandlingseffekt og hvis vi tester på signifikansniveau 5%, så skal vi i ca. 50 tilfælde få forkastet hypotesen. Den andel af testene der giver anledning til forkastelse kaldes for dækningsprocenten (eng: coverage percentage).

Hvis hypotesen forkastes f.eks. 100 gange så er dækningsprocenten 10 og det svarer til at p-værdierne er anti-konservative. Effekter forekommer at være mere signifikante end de i virkeligheden er; dvs. vi kommer til fejlagtigt at drage "for stærke" konklusioner. Tabel 7 viser resultaterne for de forskellige modeltyper.

Tabel 7: Dækningsprocenter for forskellige signifikansniveauer. De tre rækker, der markerede giver i praksis de korrekte dækningsprocenter og opnås når der tages højde for usikkerheden på variansparametrene.

	0.01	0.05	0.10
lm+F	0.21	0.31	0.41
lm+X2	0.24	0.35	0.42
avg_lm+F	0.01	0.06	0.11
avg_lm+X2	0.07	0.13	0.19
mixed+X2	0.05	0.14	0.23
mixed+F	0.01	0.06	0.11
mixed+PB	0.01	0.05	0.10

Konklusionerne er som følger:

1. Hvis man holder sig indefor den verden, der hedder lineære normale modeller så får man den rette dækningsprocent når 1) analyserer gennemsnittene og 2) baserer testene på at der tages højde for, at residualvariationen er estimeret fra data (dvs. man lave F-test i stedet for χ^2 test).

2. Hvis man betragter mixed models så er konklusionen den samme: Hvis man tager højde for at variansparametrene er estimerede fra data (og derfor laver F-test baseret på Kenward-Roger eller test baseret på parametrisk bootstrap) så får man den rette dækningsprocent. Baserer man testene på χ^2 approksimationen får man alt for store dækningsprocenter svarende til at en effekt kommer til at se mere signifikant ud end den er.

Man kan, med en hvis ret, argumetere for at de problemstillinger med tests man i de foregående afsnit har forsøgt at håndtere alle er knyttet til, at der er tale om et meget lille studium: To behandliger, tre individer per behandling og tre målinger per individ. Havde man blot haft flere individer ville problemerne forsvinde af sig selv. Men ofte er der naturlige begrænsninger på antallet af individer. Et eksempel herpå er givet i Afsnit 5.

5 Eksempel: Sukkerroer - et split plot eksperiment / hierarkisk design

Man ønsker at modellere hvordan sukkerindhold (pct) i sukkerroer afhænger af så- og høsttidspunkt. Der er fem såtidspunkter (s) og to høsttidspunkter (h). Forsøget var udlagt i tre blokke. Data findes i `pbkrtest` pakken og stammer fra et forsøg lavet ved det tidligere Danmarks JordbrugsForskning, der i dag er en del af Aarhus Universitet. I dette afsnit illustrerer vi desuden hrugen af `pbkrtest` pakken.

Forsøgsplanen er som følger:

```
# Plot allocation:
#       | Block 1       | Block 2       | Block 3       |
#       +---------------|---------------|---------------+
# Plot  | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | Harvest time
# 1-15  | s3 s4 s5 s2 s1 | s3 s2 s4 s5 s1 | s5 s2 s3 s4 s1 | Sowing time
#       |---------------|---------------|---------------|
# Plot  | h2 h2 h2 h2 h2 | h1 h1 h1 h1 h1 | h2 h2 h2 h2 h2 | Harvest time
# 16-30 | s2 s1 s5 s4 s3 | s4 s1 s3 s2 s5 | s1 s4 s3 s2 s5 | Sowing time
#       +---------------|---------------|---------------+
```

De første observationer ses i Tabel 8.

Tabel 8: De første observationer i 'beets' datasættet.

harvest	block	sow	yield	sugpct
harv1	block1	sow3	128	17.1
harv1	block1	sow4	118	16.9
harv1	block1	sow5	95	16.6
harv1	block1	sow2	131	17.0

Uanset om man betragter udbytte eller sukkerprocent viser et plot (ikke gengivet her) at der ikke er indikation af interaktion mellem så- og høsttidspunktet. En model for forsøget kunne derfor være

$$y_{hbs} = \mu + \alpha_h + \beta_b + \gamma_s + U_{hb} + \epsilon_{hbs}, \tag{1}$$

hvor $U_{hb} \sim N(0, \omega^2)$ og $\epsilon_{hbs} \sim N(0, \sigma^2)$. Bemærk at U_{hb} beskriver den tilfældige variation mellem plots (indenfor blokke). Med `lmer()` funktionen fra `lme4` pakken kan vi teste for ingen effekt af så- og høsttidspunkt som følger:

```
beet.lg <- lmer(sugpct ~ block + sow + harvest +
                    (1 | block:harvest), data=beets, REML=FALSE)
beet.noh <- update(beet.lg, .~. - harvest) # Fjern høsttidspunkt
beet.nos <- update(beet.lg, .~. - sow)     # Fjern såtidspunkt
anova(beet.lg, beet.noh)
```

```
## Data: beets
## Models:
## beet.noh: sugpct ~ block + sow + (1 | block:harvest)
## beet.lg: sugpct ~ block + sow + harvest + (1 | block:harvest)
##          Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## beet.noh  9 -69.1 -56.5   43.5    -87.1
## beet.lg  10 -80.0 -66.0   50.0   -100.0  12.9      1    0.00033
```

```
anova(beet.lg, beet.nos)
```

```
## Data: beets
## Models:
## beet.nos: sugpct ~ block + harvest + (1 | block:harvest)
## beet.lg: sugpct ~ block + sow + harvest + (1 | block:harvest)
##          Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## beet.nos  6  -2.8   5.6    7.4    -14.8
## beet.lg  10 -80.0 -66.0   50.0   -100.0  85.2      4     <2e-16
```

Begge effekter forekommer at være stærkt signifikante, men det interessante er her at sammenligne med resultaterne med Kenward-Roger og parametrisk bootstrap metoden. For såtidspunktet får man stadig meget små p-værdier, men for høsttidspunktet bliver billedet et andet.

```
KRmodcomp(beet.lg, beet.noh)
```

```
## F-test with Kenward-Roger approximation; time: 0.17 sec
## large : sugpct ~ block + sow + harvest + (1 | block:harvest)
## small : sugpct ~ block + sow + (1 | block:harvest)
##       stat  ndf  ddf F.scaling p.value
## Ftest 15.2  1.0  2.0         1    0.06
```

```
PBmodcomp(beet.lg, beet.noh)
```

```
## Bootstrap test; time: 6.48 sec;samples: 1000; extremes: 27;
## large : sugpct ~ block + sow + harvest + (1 | block:harvest)
## small : sugpct ~ block + sow + (1 | block:harvest)
##        stat df p.value
## LRT    12.9  1 0.00033
## PBtest 12.9    0.02797
```

Afslutningsvist bemærkes det, at da designet er balanceret kan man lave F-tests indenfor strata som vist nedenfor. Bemærk: F-teststørrelsen er $F_{1,2}$ for høsttidspunkt og $F_{4,20}$ for sådidspunkt.

```
beets$bh <- with(beets, interaction(block, harvest))
summary(aov(sugpct ~ block + sow + harvest +
              Error(bh), data=beets))
```

```
##
## Error: bh
##           Df Sum Sq Mean Sq F value Pr(>F)
## block      2 0.0327  0.0163    2.58   0.28
## harvest    1 0.0963  0.0963   15.21   0.06
## Residuals  2 0.0127  0.0063
##
## Error: Within
##           Df Sum Sq Mean Sq F value  Pr(>F)
## sow        4   1.01  0.2525     101 5.7e-13
## Residuals 20   0.05  0.0025
```

6 Diskussion og afsluttende bemærkninger

Eksemplerne der er vist ovenfor er sådan, at man kan komme udenom problemet med korrelerede målinger ved at beregne passende gennemsnit og analysere disse. Dette er gjort for at vise, at de metoder fra `pbkrtest` der illustreres giver de "rette svar". Den virkelige styrke ligger dog i, at man kan arbejde med generelle mixed models og stadig beregne bedre referencefordelinger for teststørrelserne og dermed få mere retvisende konklusioner.

Det noteres, at der i beregningerne i Kenward-Rogers metode er brug for at udregne $G_j\Sigma^{-1}G_j$, hvor $\Sigma = \sum_i \sigma_i G_i$ og heri er σ_i'erne ukendte parametre og G_i'erne er kendte matricer. Det kan være både tids- og pladskrævende at beregne ovenstående sum. Et alternativ for lineære mixed models er en Satterthwaite-type approksimation; denne er hurtigere at beregne og er på vej i en kommende udgave af `pbkrtest`. Et alternativ (der også virker for generaliserede lineære mixed models) er at beregne p-værdier ved parametrisk bootstrap. Slutteligt skal det nævnes, 1) at `pbkrtest` er tilgængelig på https://cran.r-project.org/package=pbkrtest, 2) at `pbkrtest` er beskrevet i Halekoh and Højsgaard (2014) og 3) at udviklingsversioner af `pbkrtest` er tilgængelige på github og kan installeres med `devtools::install_github(hojsgaard/pbkrtest)`.

Referencer

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Halekoh, Ulrich, and Søren Højsgaard. 2014. "A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – the R Package pbkrtest." *Journal of Statistical Software* 59 (9): 1–30. http://www.jstatsoft.org/v59/i09/.

Kenward, Michael G., and James H. Roger. 1997. "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics* 53 (3): 983–97.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Outlier Detection in Categorical Data

Mads Lindskou

December 12, 2019

1 Introduction

For high-dimensional categorical data, it is difficult to determine anomalies such as outliers. We describe the novel method by Lindskou et al. [2019] that uses the class of decomposable graphical models to model the relationship among the categorical variables of interest, which can be depicted by an undirected graph called the *interaction graph*. Given an interaction graph, an exact test statistic for outlier detection is then available.

An outlier is a case-specific unit since it may be interpreted as natural extreme noise in some applications, whereas in other applications it may be the most interesting observation. A universal definition of an outlier is given by Hawkins [1980]: "an observation which deviates so much from the other observations in the data-set as to arouse suspicions that it was generated by a different mechanism". The method captures this perception of an outlier using the likelihood ratio principle. In the following section we motivate the usage of this new method through an example data set which we subsequently analyze with the R package `molic` [Lindskou, 2019].

2 Motivating Example

It is a difficult task to diagnose erythemato-squamous (ES) skin diseases. Erythema is redness of the skin and squamous cells are flat thin cells that look like fish scales. ES diseases are also a major cause of skin cancer, and it is important to find robust models helping the medical experts to diagnose precisely, accurately, and inexpensively. Usually a biopsy is necessary for the diagnosis, but unfortunately, these diseases share many histopathological features as well. Another difficulty for the differential diagnosis

is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages.

Many researchers have used the UCI `dermatitis` data [Ilter and Altay Güvenir, 1998] in order to develop accurate and robust models to classify the correct ES disease of an individual. The `dermatitis` data contains 358 (8 observations removed due to missing values) observations and 34 explanatory variables where 12 are clinical attributes (erythema, itching, scaling etc.) named $c_1, c_2, \ldots, c_{12}$ and 21 histopathological attributes (melanin incontinence, spongiosis etc.) which we name $h_1, h_2, \ldots, h_{21}$. All explanatory variables take values in $\{1,2,3,4\}$ except for c_{11} (family history) which is binary and c_{12} (age) which we have discretized into six equidistant intervals. The class attribute which we name ES has the following values

- psoriasis (111 instances)
- seborrheic dermatitis (60 instances)
- lichen planus (71 instances)
- pityriasis rosea (48 instances)
- chronic dermatitis (48 instances)
- pityriasis rubra pilaris (20 instances).

Many of the classical machine learning algorithms like decision trees, random forests, nearest neighbor, support vector machines, neural networks etc. have been applied to `dermatitis` [Liu et al., 2015]. They all achieve a prediction accuracy above 95% (some even above 99%). An obvious problem with classification is that of misclassification; especially when an individual are almost equally likely to be classified as one of several different classes. This can be problematic in the `dermatitis` data since the diseases, unfortunately, share features with very little differences [Güvenir et al., 1998].

Another serious concern about classification arises in situations where a data set contains exclusive classes but not exhaustive. In other words, a patient may suffer from a skin disease, possibly not discovered yet, not in the family of ES diseases. Even more unfortunate is it to classify an individual as having a skin disease if the individual does not suffer from a skin disease at all.

Given a new observation, y, we seek to test the following hypothesis:

H_1 :y has psoriasis

H_2 :y has seborrheic dermatitis

H_3 :y has lichen planus

H_4 :y has pityriasis rosea

H_5 :y has cronic dermatitis

H_6 :y has pityriasis rubra pilaris.

It is important to note, that all tests are mutually exclusive by definition and a correction for multiple hypothesis testing is therefore not required. It is possible that more than two hypothesis can not be rejected and further investigations are then needed.

3 The Outlier Test

The outlier method takes into account the mutual dependencies between all explanatory variables using the family of decomposable graphical models. Graphical models is a family of statistical models for which the dependencies between variables can be depicted by a graph (here an undirected graph) called the interaction graph. An interaction graph consists of nodes and edges where nodes represents variables and an edge represent an association (or partial correlation) between the nodes that it connects. The term 'decomposable' refers to a subclass of graphical models for which the interaction graph is not allowed to contain cycles of length greater than four without a chord. This ensures a closed form solution of the test statistic described shortly. In Figure 1 an interaction graph with variables a, b, c, d, e, f and g is depicted. It is seen that d is directly associated with b, c, e and f, hence these cannot be assumed to be independent. Notice, that a is only connected to c through b; we interpret this as "a is independent of c when we know the value of b". Such statements can be used to gain insight into complex systems. This phenomena is also known as *conditional independence*.

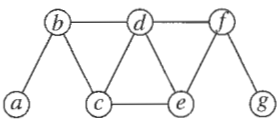


Figure 1: Example of a decomposable graph.

Given an interaction graph for a homogeneous class, k, of observations (e.g. psoriasis

patients where $k = 1$) we calculate the deviance test statistic

$$D_k(y) = -2\log LR_k = -2\log \frac{\text{Likelihood for } y \text{ having disease } k}{\text{Likelihood for } y \text{ not having disease } k}.$$

Large values of D_k are critical to the null hypothesis H_k. We need to determine if the particular value $D_k(y)$ is 'too large' or not. The outlier method works as follows:

- Simulate N observations, $y_1, y_2, \ldots, y_N$
- Calculate the deviance of y_j for $j = 1, 2, \ldots, N$
- An approximation of the density of D is given as the empirical distribution of the obtained deviances
- For a new observation, y, of interest calculate $D_k(y)$
- Determine if $D_k(y)$ is larger than the critical value corresponding to a significance level α in the approximated distribution of D_k; if so, y is declared an outlier.

Since the likelihood ratios, LR_k, is less than or equal to one, the tests are all one-sided. In other words, if $D_k(y)$ is greater than or equal to some critical value, H_k is rejected.

4 Analysis using Software

We use the R package `molic` [Lindskou, 2019] to analyze the `dermatitis` data and highlight some fallacies when using classification methods. The aim is to find observations for which we cannot reject the hypothesis of being an outlier in more than just one of the ES diseases. Hence, for all observations we construct six outlier test. Let y denote an arbitrary observation. We now show how to conduct the single hypothesis test that y is an outlier in the ES psoriasis class. First construct the psoriasis database `psor` containing 112 observations (y is included under the hypothesis) and fit an interaction graph to this data as follows

```
library(molic)
g <- fit_graph(psor, q = 0) # q = 0 correspond to AIC
plot(g)
```

It is now possible to make statements on how variables describing psoriasis patients are associated. From the interaction graph in Figure 2 we can e.g. state that c_1 and c_2 are conditionally independent given c_7. We leave it to experts to decide whether or not

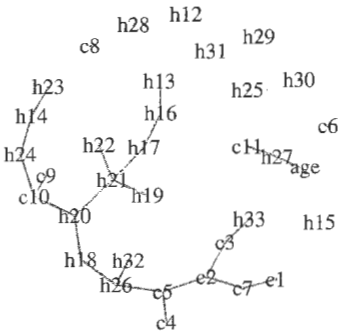


Figure 2: Interaction graph for psoriasis patients.

this has a biological meaning. The interaction graph is now used as the building block for determining if y is an outlier in `psor`:

```
m <- fit_outlier(psor, y, g)
print(m)
---------------------------------
 Simulations: 10000
 Variables: 34
 Observations: 112
 Estimated mean: 42.66
 Estimated variance: 37.55
   ---------------------------
 Critical value: 53.86771
 Deviance: 50.51582
 P-value: 0.1053
 Alpha: 0.05
 <outlier, outlier_model, list>
---------------------------------
```

Thus, we do not reject that y has psoriasis on a 5% significance level with a p-value of 0.1053. We can plot the approximated density of the deviance statistic as follows:

```
plot(m)
```

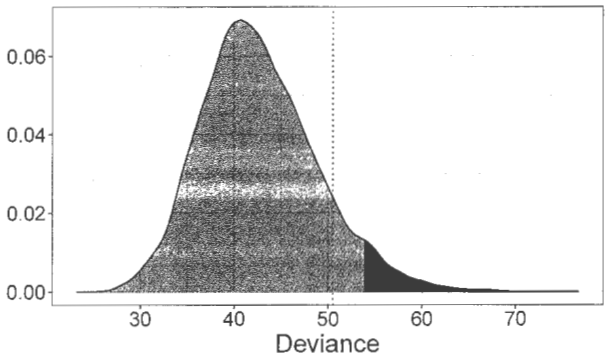


Figure 3: Approximated density of the deviance statistic in model `m`. The vertical solid line indicates the observed deviance 50.51.

The above procedure was repeated for each of the six ES diseases for all 358 observations. The number of observations declared as outliers in $2,3,4,5$ or 6 different diseases in ES is summarized in Table 1. Surprisingly, there are 5 observations that are declared as outliers in all six diseases. This should be taken seriously and further investigations would be required. As expected, most observations (305) outliers in 5 diseases and could not be rejected in the remaining one. However, there is also a non neglecting fraction, 13.4%, of observations that are outliers in more than one disease.

Number of rejected diseases:	2	3	4	5	6
Number of observations:	1	15	32	305	5

Table 1: Number of observations rejected in $2,3,4,5$ and 6 different ES diseases.

Consider now Figure 4 where all six corresponding (kernel smoothed) distributions of the deviance statistic is depicted. The dotted lines represents the observed deviance of y and the black area is the significance level (here 0.05) in the respective hypothesis. It is seen, that we cannot reject that y has either psoriasis, seboreic dermatitis or pityriasis rosea. Thus, a simple classification of y could indeed lead to a false conclusion.

```
mm <- fit_multiple_models(dermatitis, y, response = "ES")
plot(mm)
```

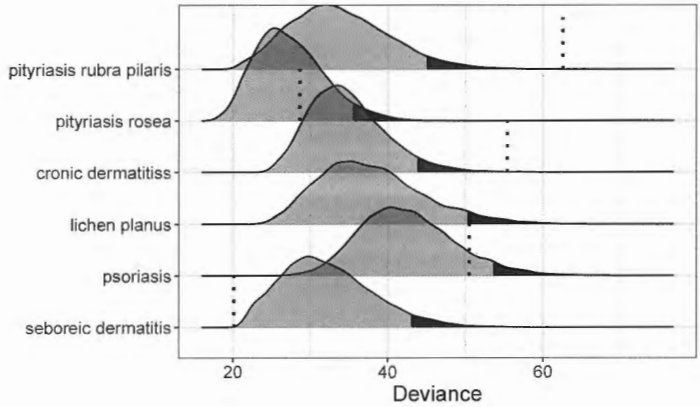


Figure 4: Six distributions of the deviance statistic. Dotted lines represents observed deviances of y and the black areas are the significance level (here 0.05). Absence of a dotted line means that it is larger than all other deviances and the hypothesis is then rejected.

References

H. A. Güvenir, G. Demiröz, and N. Ilter. Learning differential diagnosis of erythemato-squamous diseases. *Artificial intelligence in medicine*, 13(3):147–165, 1998.

D. M. Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

Ilter and H. Altay Güvenir. UCI machine learning repository, 1998. URL `https://archive.ics.uci.edu/ml/datasets/Dermatology`.

M. Lindskou. molic: An R package for multivariate outlier detection in contingency tables, Oct. 2019. URL `https://doi.org/10.21105/joss.01665`.

M. Lindskou, P. S. Eriksen, and T. Tvedebrink. Outlier detection in contingency tables using decomposable graphical models. *Scandinavian Journal of Statistics*, 2019. doi: 10.1111/sjos.12407.

T. Liu, L. Hu, C. Ma, Z.-Y. Wang, and H.-L. Chen. A fast approach for detection of erythemato-squamous diseases. *International Journal of Systems Science*, 46(5):919–931, 2015.

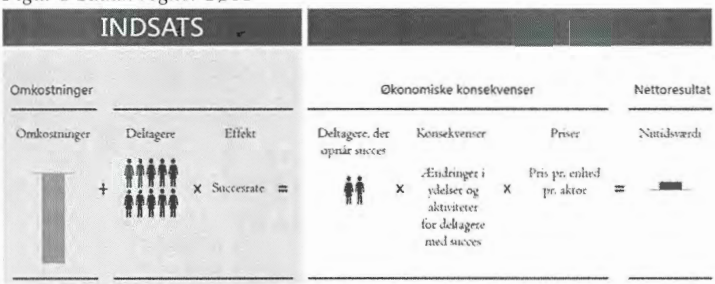
Den Socialøkonomiske Investeringsmodel - SØM

Tine Hjernø Lesner og Kenneth Lykke Sørensen, Socialstyrelsen

SØM er Socialstyrelsens beregningsværktøj, der gør det muligt at regne på de budgetøkonomiske konsekvenser over tid for de offentlige kasser ved at investere i vidensbaserede sociale indsatser. SØM er særligt målrettet kommunerne, men kan bruges af alle, der ønsker mere viden om, hvordan en social indsats påvirker den offentlige økonomi over tid. SØM er et brugervenligt værktøj i Excel, som kan downloades frit fra Socialstyrelsens hjemmeside.

Nettoresultatet for det offentlige ved en social indsats er summen af indsatsens omkostninger og de budgetøkonomiske konsekvenser, der følger af den effekt, indsatsen har på borgerne. Nedenstående figur viser, hvordan SØM regner.

Figur 1 Sådan regner SØM



SØM består overordnet af to dele:

- En beregningsramme, der beregner det samlede nettoresultat ud fra brugerens input vedrørende indsatsens omkostninger og succesrate samt de økonomiske konsekvenser over tid.
- En vidensdatabase, der indeholder viden om effekt, konsekvenser og priser for en række målgrupper på socialområdet.

Om vidensdatabasen

Vidensdatabasen indeholder viden og estimater vedrørende en række udsatte målgrupper på både voksenområdet og børne- og ungeområdet. Vidensdatabasen indeholder viden om:

1) Effekt

Vidensdatabasen indeholder effektstørrelser for forskellige sociale og psykosociale indsatser målt i danske og internationale videnskabelige studier samt rapporter fra anerkendte vidensproducenter. Der er medtaget studier, der lever op til følgende to inklusionskriterier:

- Der er målt kvantitative resultater for borgere efter deltagelse i en social indsats
- Der er sammenlignet med enten en kontrolgruppe eller resultatmålet før deltagelse i den sociale indsats for så vidt muligt at vise en kausal sammenhæng.

2) Konsekvenser

Vidensdatabasen indeholder en lang række konsekvensestimater, det vil sige bud på hvor meget forskellige offentlige ydelser og aktiviteter påvirkes, når en borger oplever effekt af en indsats. For at opgøre den mulige konsekvens af en succesfuld social indsats opdeles alle målgrupper i to adskilte grupper pba. en indikator for succes:

- en gruppe, der har opnået succes (succesgruppen)
- en gruppe, der ikke har opnået en succes (øvrige)

For fx hjemløse kan en succesfuld indsats resultere i, at en gruppe borgere ét år efter er kommet ud af hjemløshed (succes), mens den gruppe borgere, der ikke har opnået succes, fortsat vil være hjemløse. Denne "succes" ses som en approksimation for, at individerne har været i en succesfuld social indsats.

Den mulige konsekvens af en succesfuld social indsats opgøres herefter som forskellen over tid af det forbrug, succesgruppen har og det forbrug, den øvrige gruppe har, hvor der statistisk kontrolleres for sammensætningen af de to grupper og eventuel forskel i forbruget før inddelingen i succesgruppe og øvrige. En konsekvens af en succesfuld social indsats til hjemløse kan fx være et reduceret forbrug af sociale foranstaltninger efter serviceloven, reduceret kriminalitet eller større forbrug af misbrugsbehandling.

Konsekvenserne er estimeret på baggrund af danske registerdata. For hver konsekvens estimeres følgende regressionsmodel:

$$y_{it} = \alpha + \beta d_{i,t=1} + \beta y_{i,t=0} + \beta X_{i,t=0} + \mu_{i,t=0} + \varepsilon_{it} \qquad \text{for t} = 2,..,\text{N}$$

Hvor y er den økonomiske aktivitet målt i år t efter at individ i er identificeret i målgruppen. d er en indikator variabel for succesgruppe, som har værdien 1, hvis barnet i år t=1 placeres i succesgruppen. $y_{i,t=0}$ er forbruget af den økonomiske aktivitet i år $t=0$, $X_{i,t=0}$ er en vektor af baggrundsvariable om individ i målt i år $t=0$, , $\mu_{i,t=0}$ er en år fixed effect[1] og ε_{it} er et individ-specifikt fejlled. N er det samlede antal år, som individ i kan følges i registerdata. Regressionsmodellen er en OLS-model, hvor der korrigeres for robuste standardfejl. Der anvendes en Bonferroni-korrektion af 5% signifikansniveauet, der justerer for multiple hypoteser.

Konsekvenserne er naturligvis kun vejledende for den konkrete lokale kontekst, som brugeren af SØM laver sin beregning i. Brugeren har derfor mulighed for at indtaste egne bud baseret på egne data i Excel-redskabet.

3) Priser

For hver konsekvens findes der i vidensdatabasen en pris, der afspejler den værdi, konsekvenserne har for henholdsvis kommune, region og stat. Priserne anvendes til at omregne konsekvensestimaterne (opgjort som ændringer i mængden af ydelser og aktiviteter) til økonomiske konsekvenser (opgjort i kroner). Der er to centrale krav til enhedspriserne i SØM:

- Prisen skal dække omkostningen ved, at en borger får en bestemt ydelse eller offentlig service. Den må således ikke være "forurenet" af også at dække andre ydelser.
- Så vidt muligt skal alle omkostninger ved den pågældende indsats være inkluderet, dvs. også administration, ledelse, sagsbehandling og andre overheadomkostninger.

Refusion mellem stat, kommune og region er medregnet for de ydelser, hvor der er en lovgivnings- og aftalebestemt fast refusionssats. I det tilfælde bliver enhedsprisen herefter fordelt mellem stat, region og kommune.

[1] Denne tager højde for systematiske forskelle over tid. For eksempel kan der være forskel på diagnosticerings-/visitationspraksis over år, hvilket vil have betydning for hvem, der er i en målgruppe, eller hvilken indsats de tilbydes.

Priserne opgøres målgruppespecifikt, hvor data tillader det (fx sygesikringskontakter, sygehusbehandling og skat). Priserne beregnes som gennemsnitspriser. Priserne er vejledende og brugeren har mulighed for at indtaste egne bud på priser i Excel-værktøjet, hvis priserne i brugeren lokale kontekst afviger fra SØMs priser.

SØM kan give et bedre grundlag for dialog og lokalpolitiske diskussioner, men kan aldrig stå alene, når der skal træffes politiske beslutninger om investeringer i sociale indsatser. Sociale indsatser skal i høj grad også vurderes ud fra den progression og livskvalitet, der skabes for borgerne.

Sexual crime against children with disabilities: a nationwide prospective birth cohort-study

Mogens Nygaard Christoffersen, senior researcher emeritus VIVE

Abstract

Background: EU member states have ratified the Convention on the Rights of Persons with Disabilities and the Conventions of Rights of the Child. Numerous studies have shown that the rate of sexual victimization against children with disabilities is higher than the rate for children without disabilities.

Objective: The study focuses on examining sexual crime against children with disabilities and explaining differences in victimization in order to elucidate to what extent types of disability, family disadvantages, gender, high-risk behavior, location influences adolescents' risk of sexual victimization. Previous population studies lack scientifically sound research methodology and results are weak or inconclusive.

Methods: data is based on a national study of reported sexual crime against children in Denmark aged between 7 and 18 years of age using total birth cohorts (N=679,683). The statistical analysis is a discrete time Cox-model. An extended list of potential risk factors was included in the analysis in order to adjust for confounding. The potentially confounding risk-factors were collected independently from various population-based registers, e.g. employment statistics, housing statistics, education statistics, income compensation benefits, and population statistics (e.g. gender, age, location). Hospital records with information on types of disability based on the national inpatient register and national psychiatric register were collected independently of the collection of law enforcements records about reported sexual offences under the Danish Central Crime Register.

Results: Among total birth cohorts 8,039 persons or 1.18 pct. were victim of a reported sexual crime once or several times. Children with intellectual disabilities were more likely to be victimized of a reported sexual crime than non-disabled children were: ADHD odds ratio: 3.7 (3.5-3.9), mental retardation: 3.8 (3.6-4.0), autism 3.8 (3.6-4.0). This contrast with children with speech disability, stuttering and dyslexia who were less likely to be victimized when adjusted for family vulnerability and other confounding risk factors.

Conclusions: The present study finds that intellectual disability and family vulnerability e.g. parental substance abuse, parental violence, family separation, the child in care, and parental unemployment indicate an increased risk of being a victim of a sexual crime, while speech disability seems to be ensuring protection.

Assessment of risk factors may permit professionals to facilitate prevention and treatment interventions. The study underreports the size of the problem probably because adolescents with disabilities face barriers when reporting victimization.

Keywords

Sexual assault, ADHD, autism, family risk, longitudinal, official records, victim blame

1. *Introduction*

Many studies of sexual violence against children with disabilities are based on small-scale studies. However, many of the large-scale population studies lack scientifically sound research design and methodology (Leeb et al., 2012). There are poor standards of measurement of disability and insufficient assessment in the studies whether sexual abuse preceded the development of disabilities. Longitudinal panel data are rarely available. Previous studies of sexual crime against children have mainly been based on self-reported victim surveys or studies of offender population, which also suffer from risks of selection biases. These gaps need to be addressed through a whole population sample with standardized measures of disability and sexual violence (L. Jones et al., 2012).

The insufficient knowledge of causes and settings of sexual violence hampers the development of initiatives for preventing violence against children with disabilities. Information about risk factors the years preceding a crime incident against children with disabilities is limited in crime victimization surveys and existing databases. Furthermore, risk of estimates might overestimate the association between violence and disability because of inadequate adjustment for confounding (L. Jones et al., 2012; Petersilia, 2001).

There is a critical need for a theory-driven research with assess to criminal justice record, medical records, and reliably family and community risk variables in order to identify associations and pathways between child disabilities and violence against children as an important step toward planning targeted and appropriate prevention and intervention activities (Leeb et al., 2012; Sullivan, 2009).

2. *Methods*

2.1 Research design

We want to explore the prevalence of sexual offences against children with disabilities, and identify risk and protective factors for, and underlying causes of, violence against a child with disabilities. The study is based on longitudinal panel data including the whole population sample (Christoffersen, 2019). Data consists of administrative records with standardized measures of disability and sexual violence. The study examines the risk factors proceeding the first time sexual offending against a child.

2.2. Study population

Eleven national birth cohorts of children born 1984-1994 age 7-18 are followed (N=679,683) and parental and familial risk factors of victimization during adolescence are included. The possibility of

describing victims is possible after 2001. In the present study, we are looking into the window from 2001 to 2012 where victims are tracked (Figure 1).

All other data, which are indicating disadvantage during adolescence, factors associated with the person, and options of high-risk groups, current situations and possibilities, location or neighborhood are available from 1980 and forward for both children and their parents.

2.3 Data

The nationwide registers used are the following: Population Statistics, Medical Register on Vital Statistics, Causes of Death Register, Population and Housing Census, Unemployment Statistics, Education Statistics, Social Assistance Act Statistics, Income Compensation Benefits, Labour Market Research, Fertility Research, Criminal Statistic Register, National Patient Register, Danish Psychiatric Nationwide Case Register and Medical Birth Register. Professional agencies decide to incorporate data into the files based on established criteria and manualised decisions. Data are registered prospectively, and assumed to be collected independently from various numbers of agencies.

2.4 Analysis

The purpose of the present analysis is to locate relevant risk factors, such as background factors, and type of disability. We will describe both the strength (odds ratio) of different risk factors, and the overall exposure of risk factors in the population. These two components decide the risk factors' contribution to the number of victimized persons, and attributable fractions (AF) are calculated (Greenland, 2008). AF express the reduction in incidence of reported sexual violence that would be achieved if the population had not been exposed at all compared with the current exposure pattern (Greenland & Drescher, 1993).

The data is analyzed by the discrete time-Cox-model (Allison, 1982). A discrete-time model 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). An event is a police report on a sexual crime. Individuals' event history is broken up into 12 set of discrete time-units (age 7 to 18 years) in which an event either did or did not occur (Christoffersen, 2019).

Each individual is observed until 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. Consequently, individuals are excluded from the case group and controls after the first event. The person-years at risk were constructed for the total birth cohorts. Pooling the non-censored years of all individuals, the person-years, made the numbers at risk (N= 4,342,561).

The discrete-time Cox model is used to allow for changing covariates over time. The risk factors and measures of disability are divided into three types for the purpose of this study. The Type I risk factors are those that are taken to be indicative throughout the risk period, irrespective of the year when the risk factor was notified. Risk factors of Type II, in contrast, identify the presence of that factor in the year prior to the event. Finally, the Type III risk factors act on the following year and all the subsequent years when observed the first time (Christoffersen, 2019).

2.4.2 Ethical considerations

We have no information whether an alleged perpetrator could be a caregiver, peers or others. A limited number of the alleged perpetrators can be identified with the help of the Danish Central Crime Register, but we have chosen not to include information about the alleged perpetrator in the present study in order to avoid selection bias. The study only contain information about the victims, their families and only few structural risk factors, while background factors about the perpetrator are missing. The data structure therefore impose a risk of blaming the victim as a superficial examination of results.

2.5 Measures

2.5.1 Sexual crime (dependent variable)

Article 19 of the Convention on the Rights of the Child (CRC): violence is understood as any form of physical or mental violence, injury, abuse, neglect or negligent treatment, maltreatment, or exploitation, including sexual abuse (CRC,Committee on the Rights of the Child, 2011). In the present study, we have for practical reasons narrowed the definition of violence to sexual offences i.e. rape, sexual assault, sexual exploitation, incest, indecent exposure. Sexual assault involving sexual intercourse/penetration without consent or defined unlawful because of the age of the victim and or/ the relationship between the victim and the offender. Sexual crimes are reported criminal offences against

the person according to law enforcement records. The crime data are collected from the criminal records, which include children who have been victims of sexual crimes within or without the family.

Table 1 Classification of indicators of disabilities.

Disabilities:	**International Statistical Classification of Diseases and related Health Problems ICD-10**
Autistic spectrum disorder	Autism F84 Pervasive developmental disorders
Speech disability	Specific developmental disorders of speech and language ICD-10:F80, Infantile cerebral palsy ICD-10: G80, Speech disturbances ICD-10:R47, Lack of expected normal physiological development ICD-10:R62
ADHD	Diagnosed with ADHD in a psychiatric ward according to the Danish Psychiatric Nationwide Case Register. ADHD F90 Hyperkinetic disorders and/or receiving ADHD-drugs 'N06BA04' or 'N06BA09'
Loss of hearing	Conductive and sensorineural hearing loss ICD-10:H90-H91, Other disorders of ear ICD-10:H93-95
Epilepsy	Acquired aphasia with epilepsy ICD-10:F80.3, Epilepsy ICD-10:G40
Mental retardation	Mental retardation ICD-10:F70-F79
Down's syndrome	Chromosomal abnormalities not elsewhere classified: ICD-10: Q90
Brain injury	Intracranial injury S06, Other mental disorders due to brain damage and dysfunction and to physical disease F06, Post-concussion syndrome ICD-10: F07.2, Chronic post-traumatic headache ICD-10: G44.3
Stuttering	Stuttering and other behavioral and emotional disorders with onset usually occurring in childhood and adolescence ICD-10:F98
Physical disabilities	Symptoms and signs involving the nervous and musculoskeletal systems ICD-10:R25-R29, Injuries of neck and trunk, limp, ICD-10:T91-T94
Dyslexia	Specific developmental disorders of scholastic skills ICD-10:F81. Dyslexia and other dysfunctions, not elsewhere classified ICD-10:R47
Blindness	Low vision ICD-10:H54
Congenital malformations	Congenital malformations, deformations and chromosomal abnormalities ICD-10:Q00-99

Note an example: ICD-10 diagnoses 'F84' includes all diagnoses F84.0-F84.9

Source: (World Health Organization 1992)

2.5.2 Disabilities (independent variable)

Some of the disabilities are severe, life-long disabilities attributable to mental and/or physical impairments, manifested before 18 years of age (Type I). Examples of this are the autistic spectrum of disorders, ADHD, mental retardation, epilepsy, sensory impairment (e.g. loss of hearing, blindness), Down's syndrome, and other congenital malformations.

Some disabilities are assumed to be acquired (Type III) and are therefore only recorded when found. These include speech disability (e.g. developmental disorders of speech and language, cerebral palsy), brain injury (e.g. mental disorders due to brain damage and dysfunction, post-concussion syndrome, chronic post-traumatic headache), stuttering (e.g. behavioral and emotional disorders with onset usually occurring in childhood and adolescence), dyslexia, and physical disabilities (i.e. orthopedic impairment).

We classified disabilities into 13 main groups, which did not cover all disabilities (Table 1). The categories did not include disabilities, which could be consequences of maltreatment such as internalizing disorders, depression, anxiety, post-traumatic disorder and other emotional disorders.

2.5.3 Risk factors (covariates)

Parental background factors such as parental violence (domestic and otherwise), parental inpatient mental illness, parental suicidal behavior or alcohol abuse, parental long-term unemployment, family separation, child in (public) care, are included into the regression analysis. Structural factors such as the victim living in a disadvantaged area, and non-Danish citizens are included. Substance abuse indicated risk-taking behavior as a precursor of becoming a victim of sexual violence during adolescence.

3. Results

The population followed (N=679,683) comprised 8,039 or 1.18 % victims of a sexual assault during 2001 to 2012 when they were between 7 and 18 years old. The number of person-years were 4,342,561.

Table 2 presents differences in victimization rates between person-years with and without various disabilities. The most common disability was ADHD. Ten percent, of the person-years, the adolescent suffered from ADHD, while eight or nine percent the individual suffered from autistic spectrum disorders, while eight percent of the person-years the child suffered from mental retardation.

TABLE 2 INDICATORS OF DISABILITY

Odds ratio for types of disability prior to first-time victim of a sexual crime.

Person-years for children born in 1984-1994 (age 7 to 18 years old). Bivariate results from a discrete time Cox analysis.

Factors associated with disability:	Type	% of controls	% of Cases	OR	95% CI
Autistic spectrum disorders	(I)	8.9	27.0	3.8	[3.6-4.0]
Speech disability (i.e. cerebral palsy)	(III)	8.2	10.2	1.3	[1.2-1.4]
ADHD	(I)	10.8	30.8	3.7	[3.5-3.9]
Loss of hearing	(I)	1.1	1.8	1.6	[1.4-1.9]
Epilepsy	(I)	1.6	2.1	1.9	[1.6-2.1]
Mental retardation	(I)	8.0	24.7	3.8	[3.6-4.0]
Down's syndrome	(I)	0.1	-	Ns	
Brain injury	(III)	5.7	8.1	1.5	[1.3-1.6]
Stuttering	(III)	2.8	5.9	2.1	[1.9-2.3]
Physical disabilities (i.e. orthopedic impairment)	(III)	1.3	1.6	1.2	[1.0-1.5]
Dyslexia	(III)	1.4	2.7	1.9	[1.7-2.2]
Blindness	(I)	0.06	0.12	2.0	[1.1-3.7]
Congenital malformations	(I)	0.4	0.6	Ns	

Note:

Some of the victims had multiple disabilities.

Type of time-dependency

Type I: disability factor observed at time t also covers the years before and after the years under investigation.

Type II: exposed to risk factor at time t then the risk factors is also present at t+1.

Type III: exposed to disability factor at time t then risk factor is also present at all the following years.

Ns means: non-significant; OR = odds ratio; 95% CI = confidence interval.

Associations with victimization vary across different disabilities. Children's person-years with disability had significantly higher rates of victimization relative to children with no disability. Exception from this pattern were Down's syndrome and other congenital malformations probably because of very few observations (approximately 0.5 % of the surveyed person-years). Table 2 shows that 27% among the victims were adolescents with Autistic spectrum disorders, 30% among the victims were adolescents with ADHD, while 24% of the victims were adolescents with mental retardation. These figures correspond to odds ratio of 3.7 to 3.8 (95% CI= 3.5-4.0). Speech disability were found among 10% of the victims, which correspond to odds ratio of 1.3 (95% CI=1.2-1.4). Some of the victims had multiple disabilities.

TABLE 3.1 FAMILY VULNERABILITY

Odds ratio for risk factors prior to first-time victim of a sexual crime.

Person-years for children born in 1984-1994 (age 7 to 18 years old). Adjusted results from a discrete time Cox analysis. (continued)

Risk factors:	Type	% of controls	% of cases	OR	95% CI	AF %
Disadvantages during adolescence						
Parental suicidal behavior	(I)	10.4	21.4	Ns		-
Mother mental retardation	(I)	4.8	9.9	Ns		-
Father mental retardation	(I)	3.9	8.4	Ns		-
Parental inpatient mental illness	(I)	16.9	31.9	Ns		-
Parental substance abuse	(I)	17.4	34.2	1.4	[1.3-1.5]	6.5
Parent diagnosed with ADHD	(I)	8.4	14.8	Ns		-
Parental violence	(III)	13.9	28.5	1.4	[1.3-1.4]	5.3
Child ever in care	(III)	2.1	7.9	1.9	[1.8-2.1]	1.9
Family separation	(III)	36.2	59.5	1.8	[1.7-1.9]	22.5
Mother teenager	(I)	2.0	5.1	1.8	[1.6-2.0]	1.6
Parent unemployed > 21 weeks	(II)	7.0	11.0	1.3	[1.2-1.4]	2.1

Table 3.1 shows that family vulnerability such as parental substance abuse, parental violence, teenage motherhood, child in public care, family separation and long-term unemployment predicted an elevated risk of victimization of a sexual assault. Odds ratio were found between 1.2 and 2.0, when accounted for multiple disabilities, and other risk factors. The attributable fractions were estimated to be 6.5%, 5.3% and 22.5% for parental substance abuse, parental violence and family separations, respectively.

Table 3.2 shows that females were between eight and nine times more at risk for being a victim of a sexual assault (95% CI=8.2-9.4). Adolescents with an alcohol abuse situation the risk of sexual assault were significantly higher (Odds ratio 1.6; 95% CI=1.4-1.9) when accounted for other risk variables.

Only few of the control person-years were exposed to alcohol abuse. On the contrary being a member of an ethnic minority seems to be a protective factor. The adjusted odds ratio were estimated to 0.7 (95% CI=0.6-0.7) for non-Danish person-years of school age children. The individual vulnerability apart from the person-years with disabilities could not explain the variations of victimizations.

TABLE 3.2 INDIVIDUAL VULNERABILITY, DISADVANTAGED AREA AND MINORITY

Odds ratio for risk factors prior to first-time victim of a sexual crime.

Person-years for children born in 1984-1994 (age 7 to 18 years old). Adjusted results from a discrete time Cox analysis. (continued)

Risk factors:	Type	% of controls	% of cases	OR	95% CI	AF %
Adolescent is female[1]	(I)	48.3	89.1	8.8	[8.2-9.4]	-
High risk behavior:						
Suicide attempt	(III)	2.0	4.6	Ns		-
Drug abuse	(II)	0.6	2.3	Ns		-
Alcohol abuse	(II)	0.9	3.1	1.6	[1.4-1.9]	0.5
Convicted of violence	(III)	0.4	0.3	Ns		-
Location or neighborhoods:						
Disadvantaged area	(II)	2.2	2.6	Ns		-
Non-Danish	(II)	8.5	5.4	0.7	[0.6-0.7]	-

1 The AF (i.e. attributable fraction) is not calculated because the risk factor is not changeable.

TABLE 3.3 INDICATORS OF DISABILITY

Odds ratio for risk factors prior to first-time victim of a sexual crime.

Person-years for children born in 1984-1994 (age 7 to 18 years old). Adjusted results from a discrete time Cox analysis.

Risk factors:	Type	% of controls	% of cases	OR	95% CI	AF %
Factors associated with disability:						
Autistic spectrum disorders	(I)	8.9	27.0	1.4	[1.2-1.5]	3.4
Speech disability	(III)	8.2	10.2	0.9	[0.8-0.9]	-
ADHD	(I)	10.8	30.8	1.9	[1.7-2.1]	8.9
Loss of hearing	(I)	1.1	1.8	1.4	[1.1-1.6]	0.4
Epilepsy	(I)	1.6	2.9	1.4	[1.2-1.5]	0.6
Mental retardation	(I)	8.0	18.6	1.2	[1.1-1.3]	1.6
Down's syndrome	(I)	0.1	-	Ns		-
Brain injury	(III)	5.7	8.1	1.1	[1.0-1.2]	0.6
Stuttering	(III)	2.8	5.9	0.8	[0.7-0.9]	-
Physical disabilities (i.e. orthopedic impairment)	(III)	1.3	1.6	1.2	[1.0-1.5]	0.3
Dyslexia	(III)	1.4	2.7	0.8	[0.7-1.0]	-
Blindness	(I)	0.06	0.12	1.6	[1.1-2.4]	-
Congenital malformations	(I)	0.4	0.6	Ns		-

Tables 3.3 shows that the predicative effect sizes of intellectual disabilities on victimization were considerable reduced when adjusted for other risk factors. While the bivariate results in Table 2 shoved

odds ratio between 3.7 and 3.8 for intellectual disabilities (e.g. autism, ADHD, mental retardation) the adjusted effect size were estimated to 1.4, 1.9 and 1.2, respectively. Even when adjusted for family vulnerability and other risk factors, the attributable contribution of autism, ADHD and mental retardation were 3.4%, 8.9% and 1.6% of the cases of sexual assaults, respectively.

The bivariate results showed that speech disability, stuttering, and dyslexia were significant predictors of victimization of sexual assault during adolescence. When the confounders were taken into account, these communicative disabilities turned out to be protective factors with odds ratios significant lower than 1.0 (compare Table 2 with Table 3.3).

4. Discussion

Previous research find that children with these intellectual disabilities and mental disabilities such as depression, anxiety, posttraumatic disorder, emotional disorder or conduct disorder are 4.6 times more likely to be exposed to sexual assaults (L. Jones et al., 2012). We suppose that these figures overestimates the association between type of disability and risk of sexual assaults because the mentioned mental disabilities may very well be a reaction to child maltreatment e.g. sexual or violent victimization. We decided not to include the mental disabilities into the list of disabilities (independent risk factors). We assume that our estimate of the odds ratio between ADHD, autism or mental retardation of 3.7-3.8 (95% CI= 3.5-4.0) is a more accurate estimate, and this estimate is furthermore based on a relatively large population register data collected prospectively.

The results give supportive evidence that an attributional part of the incidences may be due to family vulnerability or a combination between intellectual disability and family vulnerability. The attributable fraction of family vulnerability such as parental substance abuse, parental violence, family separation, and long-term unemployment express the reduction in incidence of reported sexual violence that would be achieved if the population had not been exposed to family vulnerability. The estimates indicate that the link between children with intellectual disabilities and sexual assaults is considerable weakened when the regression analyses include information about family vulnerability. The link is likely to become weaker with a better functioning family. Our findings pointed to the possibility that family functioning could have critical effect on the overall increased risk between intellectual disabilities and victimization of sexual assaults consistently found in the literature.

4.2 Limitations

The study focused on the time sequence between disabilities, risk factors and victimization, on should still be cautious before drawing conclusions on the causal direction. We have tried to establish the causal direction in the relationship between disabilities and victimization in the longitudinal study, as has been recommended by several researchers (Brown, Cohen, Johnson, & Salzinger, 1998; Chan et al., 2018; L. Jones et al., 2012). Although, we have included various risk factors associated with the risk for sexual assaults, we have only examined a limited number of variables, and have therefore been unable to investigate all the possible combined effects of multiple risk factors for sexual assaults. In order to fully understand sexual assaults against children, there is an urgent need for more knowledge about the perpetrator (Carstensen, Kongstad, Larsen, & Rasmussen, 1981). The study use administrative registers, which are useful because professional agencies decide to incorporate data into the files on the basis of established criteria and manualised decisions, and the data are registered prospectively. Furthermore, data are assumed to be collected independently from various numbers of agencies, and data completely cover all the calendar years from the children's birth until early adulthood. Still, we have a suspicion that Berkson's bias may influence the data (Berkson, 1946; Schwartzbaum, Ahlbom, & Feychting, 2003). Some diagnoses might only be reported in connection with elucidation of the aetiology of other diagnoses. This could create a spurious association in hospital's records, which is not found to that extent in the population. Finally, assessment of risk factors may permit professionals to facilitate prevention and treatment interventions. Professionals must recognize that the present study probably underreports the size of the problem because adolescents with disabilities face barriers when reporting victimization. A substantial number of victims elect not to report to the police because the victims realize or hope they can minimize their losses by avoiding the legal system (Doerner & Lab, 2015).

5. Conclusion

The study revealed disability-specific longitudinal pattern of child victimization, and intellectual disabilities predicted increased risk of victimization. The study found family vulnerability as a predictor of increased risk of victimization of sexual assaults among children with and without specified disabilities.

GAILS BIAS: KURIOSUM ELLER RELEVANT FEJLKILDE?

SÖREN MÖLLER

OPEN - Open Patient data Explorative Network, Odense Universitetshospital og Klinisk Institut, Syddansk Universitet, Odense

INDLEDNING

Det er alment kendt at summen af to (2×2)-tabeller, der udviser samme association (f.eks. som odds ratio, OR), men forskellige marginale fordelinger, kan resultere i en samlet 2×2-tabel med en afvigende association. For eksempel[1]

$$\begin{bmatrix} 30 & 50 \\ 40 & 80 \end{bmatrix} \qquad \begin{bmatrix} 30 & 150 \\ 20 & 120 \end{bmatrix} \qquad \begin{bmatrix} 60 & 200 \\ 60 & 200 \end{bmatrix}$$
$$OR = 1.20 \qquad OR = 1.20 \qquad OR = 1.00$$

Dette fænomen betegnes normalt som Simpsons paradoks efter [5][2]. Mens fænomenet er velkendt for (2×2)-tabeller, får det langt mindre opmærksomhed i ikke-linære regressionsmodeller, på trods af at det mindst siden [3] har været dokumenteret at lignede effekter gør sig gældende[3].

Fænomenet virker specielt overset i sundhedsvidenskabelig forskning, hvor det er ukendt for mange epidemiologer[4], ikke nævnes i typiske lærebøger, og kun yderst sjældent indrages i videnskabelige artikler, både i forbindelse med fastlæggelsen af analysestrategier, og i diksussion af mulige begrænsninger for studiet.

Fokus i sundhedsvidenskabelig foskning (specielt epidemiologi) er typisk på konfundering, hvor estimater for associationen mellem en eksponering X og et udfald Y konfunderes af en konfunder Z, som er associeret med både X og Y. Til gengæld diskuteres der sjældent den bias, der kan opstå for estimatet af associationen mellem X og Y, som blot skyldes at Z som selvstændig prædiktor er associeret med Y, men (potentielt) er uafhængigt af X. Se dog [1] for et tilfælde hvor denne biaskilde og Simpsons paradoks diskuteres i forbindelse med en konkret sundhedsvidenskabelig problemstilling.

Dette bidrag ønsker at undersøge størrelsen (og retningen) af denne bias, for at styrke viden om, hvor vigtigt det er at være opmærksom på biaset i planlægning og gennemførsel af forskningsprojekter.

MATEMATISK BAGGRUND

I det følgende vil den stokastiske variabel Y altid betegne udfaldet, og X den eksponering vi interesserer os for, sådan at det er et mål for associationen mellem X og Y vi ønsker at beregne. Z vil betegne en selvstændig prædiktor, som er associeret med Y, men uafhængig af X. Målet for associationen vil afhænge af den specifikke regressionsmodel.

Tak til Karen Steenvinkel Pedersen (UCL) for gode kommentarer til manuskriptet.

[1] Eksemplet er en modificeret udgave af eksemplet fra [5].

[2] Selve Simpsons eksempel (og betegnelsen af *paradoks*) er dog blevet kritiseret i litteraturen [6].

[3] Denne artikel af Gail fra 1984 er årsagen til titlen på dette bidrag.

[4] Min eneste evidens for denne påstand er dog den observerede overraskelse, når fænomenet nævnes til seminarer/konferencer.

Matematisk set er problemet (her følger vi [4]) at mange modeller ikke er kollapsible, altså at de generaliserede lineære modeller (GLM)[5]

$$g(E(Y|X=x,Z=z)) = \beta_0 + \beta_X x + \beta_Z z$$

og

$$g(E(Y|X=x)) = \beta_0 + \beta_X x$$

ikke resulterer i det samme estimat for β_x, på trods af uafhængighed mellem X og Z.

Gails artikler fra 1984 [3] dokumenterer dette problem for forskellige typer GLM, af speciel interesse for sundhedsvidenskabelig forskning er her nok, at problemet opstår for logistisk regression, men ikke for linær regression. Desuden dokumenterer denne artikel også at fænomenet optræder for Cox-regression, på trods af, at denne teknisk set ikke opfylder ovenstående formler på grund af den ikke-parametriske baselinehazard. Gail og medforfattere dokumenterer størrelsen af biaset på baggrund af enkle simulationer, samt asymptotiske beregninger, herunder at biaset i langt de fleste tilfælde vil være konservativt.

Simulation

For at undersøge størrelsen af Gails bias har vi gennemført en simulation, hvor der i alle tilfælde er blevet simuleret 1000 observationer og 1000 simulationskørsler. Vi har simuleret data for lineær regression, logistisk regression, probit-regression og Cox-regression[6] for forskellige valg af β_X og β_Z og med X og Z uafhængige dikotome vaiable, med $P(X=-0.5) = P(X=0.5) = P(Z=-0.5) = P(Z=0.5) = 0.5$, henholdsvis med X og Z uafhængige og uniformt fordelte på $[-1,1]$.

Ud fra simulationsresultaterne (nedenfor) kan vi observere at biaset, som forventet ikke optræder, i lineær regression, som jf. [3] er kollapsibel. Til gengæld optræder biaset i de andre undersøgte modeller, uafhængig af fordelingen på X og Z, men dog i alle tilfælde som et konservativt bias. Biaset er stærkere, jo større β_Z er i forhold til β_X og har cirka samme størrelse for de tre ikke-lineære modeller.

Diskussion

Ud fra både Gails originale resultater og vores simulationsresultater, ses det at dette fænomen godt kan resultere i et bias stort nok, til at være af betydning for fortolkningen af den undersøgte association. Derfor bør biaset skænkes mere opmærksomhed i sundhedsvidenskabelig forskning og undervisning. Konkret, kan der opfordres til:

- At der overvejes justering for kendte, observerede selvstændige prædiktorer, med stærk association til udfaldet også i tilfælde, hvor der ikke forventes association mellem prædiktoren og eksponeringen af interesse.
- At dette også overvejes i randomiserede studier, såfremt modellen er ikke-lineær. Der er i litteraturen opfordringer, til justering i randomiserede studier, men ofte begrundes dette med styrkeovervejelser[7], frem for biasovervejelser.
- At ikke-observerede mulige stærke prædiktorer for udfaldet indrages i fortolkningen af studiers resultater og diskussion af studiers begrænsninger.
- At forskel i omfanget af justering for stærke selvstændige prædiktorer indrages i fortolkningen af sammenligningen af associationer fra forskellige studier.
- At ovenstående indrages i statistik- (og potentielt epidemiologi-)undervisningen på sundhedsvidenskabelige uddannelser.

[5] For nemheds skyld er $\beta_0 = 0$ i alle nedenstående simulationer.

[6] Med Weibull-fordelte overlevelsestider.

[7] Se [7] for et simulationsstudie og [8] for et review af justeringspraksis i randomiserede studier, samt [2] for det europæiske lægemiddelagenturs anbefalinger om justering i randomiserede lægemiddelforsøg.

Model	Fordeling X og Z	β_X	β_Z	Gns. (s) b_X ujusteret	Gns. (s) b_X justeret for Z	Abs. bias Gns. (s)	Rel. bias Gns. (s)
Lineær	dikotom	1	1	0.9497 (0.0619)	0.9976 (0.0620)	-0.0479 (0.0029)	0.9518 (0.0041)
Logistisk	dikotom	1	1	0.9531 (0.1265)	1.0149 (0.1350)	-0.0618 (0.0190)	0.9393 (0.0159)
Probit	dikotom	1	1	0.8885 (0.0762)	1.0077 (0.0865)	-0.1192 (0.0238)	0.8820 (0.0198)
Cox	dikotom	1	1	0.8583 (0.0643)	1.0043 (0.0721)	-0.1459 (0.0286)	0.8549 (0.0255)
Lineær	dikotom	1	2	0.9988 (0.0636)	1.0018 (0.0636)	-0.0030 (0.0001)	0.9970 (0.0002)
Logistisk	dikotom	1	2	0.7897 (0.1167)	1.0041 (0.1507)	-0.2145 (0.0453)	0.7871 (0.0260)
Probit	dikotom	1	2	0.6143 (0.0611)	1.0063 (0.1042)	-0.3920 (0.0577)	0.6114 (0.0294)
Cox	dikotom	1	2	0.6893 (0.0517)	1.0039 (0.0715)	-0.3145 (0.0389)	0.6870 (0.0279)
Lineær	dikotom	2	1	2.0262 (0.0653)	1.9998 (0.0653)	0.0264 (0.0017)	1.0132 (0.0009)
Logistisk	dikotom	2	1	1.9032 (0.1364)	2.0182 (0.1473)	-0.1151 (0.0353)	0.9432 (0.0160)
Probit	dikotom	2	1	1.7818 (0.0859)	2.0150 (0.1034)	-0.2332 (0.0489)	0.8847 (0.0213)
Cox	dikotom	2	1	1.7426 (0.0807)	2.0084 (0.0890)	-0.2658 (0.0374)	0.8678 (0.0170)
Lineær	dikotom	2	2	1.9982 (0.0664)	1.9987 (0.0664)	-0.0001 (0.0000)	0.9998 (0.0000)
Logistisk	dikotom	2	2	1.6123 (0.1251)	2.0163 (0.1626)	-0.4040 (0.0738)	0.8003 (0.0280)
Probit	dikotom	2	2	1.2818 (0.0732)	2.0182 (0.1383)	-0.7363 (0.1103)	0.6366 (0.0352)
Cox	dikotom	2	2	1.3503 (0.0580)	2.0075 (0.0876)	-0.6572 (0.0532)	0.6729 (0.0180)
Lineær	uniform	1	1	0.9764 (0.0534)	0.9965 (0.0534)	-0.0201 (0.0011)	0.9797 (0.0015)
Logistisk	uniform	1	1	0.9572 (0.1171)	1.0078 (0.1269)	-0.0505 (0.0174)	0.9504 (0.0142)
Probit	uniform	1	1	0.8809 (0.0720)	1.0045 (0.0855)	-0.1236 (0.0248)	0.8774 (0.0194)
Cox	uniform	1	1	0.8414 (0.0566)	1.0044 (0.0636)	-0.1629 (0.0293)	0.8380 (0.0261)
Lineær	uniform	1	2	1.0429 (0.0551)	1.0020 (0.0552)	0.0409 (0.0011)	1.0409 (0.0026)
Logistisk	uniform	1	2	0.8150 (0.1020)	1.0088 (0.1361)	-0.1938 (0.0430)	0.8093 (0.0242)
Probit	uniform	1	2	0.6402 (0.0574)	1.0092 (0.1022)	-0.3689 (0.0555)	0.6357 (0.0263)
Cox	uniform	1	2	0.6406 (0.0441)	1.0041 (0.0629)	-0.3635 (0.0389)	0.6383 (0.0275)
Lineær	uniform	2	1	2.0156 (0.0564)	1.9985 (0.0564)	0.0171 (0.0009)	1.0085 (0.0005)
Logistisk	uniform	2	1	1.9045 (0.1350)	2.0174 (0.1466)	-0.1129 (0.0337)	0.9443 (0.0151)
Probit	uniform	2	1	1.7504 (0.0939)	2.0153 (0.1167)	-0.2649 (0.0530)	0.8690 (0.0222)
Cox	uniform	2	1	1.6873 (0.0705)	2.0079 (0.0795)	-0.3205 (0.0412)	0.8405 (0.0184)
Lineær	uniform	2	2	2.0550 (0.0577)	1.9990 (0.0578)	0.0560 (0.0015)	1.0280 (0.0011)
Logistisk	uniform	2	2	1.6049 (0.1208)	2.0101 (0.1626)	-0.4052 (0.0688)	0.7991 (0.0244)
Probit	uniform	2	2	1.2718 (0.0698)	2.0221 (0.1325)	-0.7503 (0.0913)	0.6299 (0.0266)
Cox	uniform	2	2	1.2627 (0.0514)	2.0076 (0.0782)	-0.7449 (0.0535)	0.6292 (0.0183)

Referencer

[1] W. K. Chan and D. A. Redelmeier, *Simpson's Paradox and the Association Between Vitamin D Deficiency and Increased Heart Disease*, Am J Cardiol **110** (2012), 143–144.

[2] Europeau Medicines Agency, *Guideline on adjustment for baseline covariates in clinical trials* (2015).

[3] M. H. Gail, S. Wieand, and S. Piantadosi, *Biased Estimates of Treatment Effect in Randomized Experiments with Nonlinear Regressions and Omitted Covariates*, Biometrika **71** (1984), no. 3, 431–444.

[4] S. Greenland, J. M. Robins, and J. Pearl, *Confounding and Collapsibility in Causal Inference*, Statistical Science **14** (1999), no. 1, 29–46.

[5] Simpson E. H., *The Interpretation of Interaction in Contingency Tables*, J R Stat Soc Ser B **13** (1951), 238–241.

[6] M. A. Hernán, D. Clayton, and N. Keiding, *The Simpson's paradox unraveled*, International Journal of Epidemiology **40** (2011), 780–785.

[7] B. C. Kahan, V. Jairath, C. J. Doré, and T. P. Morris, *The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies*, Trials **15** (2014), no. 139.

[8] N. Saquib, J. Saquib, and J. P. A. Ioanuidis, *Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study*, BMJ **347** (2013), no. f4313.

The Pizza Margherita Index

Sara Armandi

1 Introduction

Everyone loves pizza. Even though pizza origins from Italy, the Danes has taken it in as a loved everyday dish. Especially when not in the mood for cooking, pizzas are a common go to. Hence, pizzerias are a common sight in all parts of Denmark. The local pizzeria is often the first place visited in the neighborhood when moving to a new place – this is of course due to the mandatory "gratitude moving pizza" that friends and family deserve after helping with the move.

As pizzas are widespread in Denmark, it might be interesting to use the pizza as a national measure. But what kind of measure? Each year The Economist, a weekly magazine-format newspaper with a main focus on economics, publishes The Big Mac Index. The index was invented in 1986 as a lighthearted guide to whether currencies are at their "correct" level. It is based on the theory of purchasing power parity (PPP), the notion that in the long run exchange rates should move towards the rate that would equalize the prices of an identical basket of goods and services in any two countries [The Economist, 2020]. In this case, a burger serves as an informal measure, which makes it possible to compare economic productivity and standards of living between countries. Why shouldn't a pizza be able to do the same?

In this article data on 2,193 restaurants from Just Eat, the online takeaway platform, are utilized. All restaurants are listed with addresses, and hence, can be divided into one of the 98 municipalities in Denmark. The prices of pizza Margheritas are collected for all restaurants, having this specific pizza on their menus. The average pizza Margherita price across all municipalities are compared. Finally, as an attempt to explain the price differences, background information on the municipalities are gathered, and a few machine learning models are applied.

2 Theory

Pizza Margheritas have different prices, depending on the restaurant you visit, and where in the country you are. In order to explain these differences on a municipality level, a few machine learning models are applied. SAS® Viya® makes the creation of machine

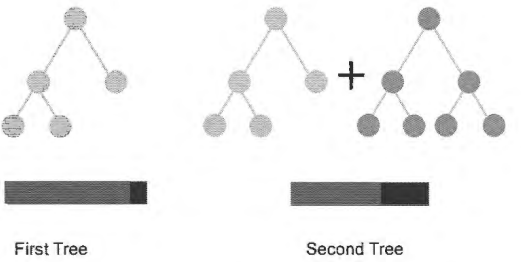
learning models very easy. In this section the decision tree and the gradient boosting model are briefly explained.

2.1 Decision Trees and Gradient Boosting

Decision Tree Models (DTM) are easy interpretable machine learning models, used for classification and regression. The name comes from the structure of the model, as data is divided into smaller subsets, while at the same time an associated decision tree is incrementally developed. When the data subset contains observations all with same characteristics, the model is complete.

The DTM consists of decision nodes and leaf nodes. In Figure 1, the first tree to the left is an example of a DTM. The first circle is called the root node and is the best predictor in the model. It represents the entire dataset. The root node is also a decision node, as it splits into further subsets. Nodes that do not split are leaf nodes and represent a final classification or decision. The tree in Figure 1 has three leaf nodes.

Figure 1: Gradient Boosting Decision Trees



Source: https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus

A Gradient Boosting Model (GBM) is an ensemble of a series of decision trees. An example of a GBM is displayed in Figure 1, where the two boxes below the trees reflects the overall error. What is important in the GBM is that the trees depend on each other, and thus, can't be created in parallel. The interpretability of the GBM is very low, as the second/latter decision tree created will not predict the same target as in the first tree. The subsequent trees try to predict how far of the original predictions were from the truth, by using the residuals from the prior trees. In this way, each tree of the GBM slowly decreases the overall error of the previous trees. Hence, the predictive power of GBM is very high, but the interpretability is low [SAS Software YouTube Channel, 2017].

3 Data

The obvious choice when in need of Danish pizza data, is to look at the homepage www.just-eat.dk. Just Eat is a worldwide online business founded in Denmark in 2001 and is the Danes' preferred online takeaway platform. The online market share of Just Eat is over 75 pct. in Denmark, and more than 13 million meals are distributed to Danes each year [Just Eat, 2020]. Using web scraping techniques, a lot of macro coding and text functions in SAS, data on 2,193 restaurants with unique restaurant ID's has been collected.

3.1 Just Eat Data

The initial data extraction was conducted on November 25th, 2019, 10:37-10:47am. The data includes information from the general webpage: https://www.just-eat.dk/takeaway/alle-byer/ which contains a list of 472 cities. By selecting each city, and each of the up to five different subpages, the dataset contains 10,097 observations, each representing a restaurant[1]. The information from each restaurant include name, address, number of ratings, type of food as well as an URL to the restaurant menu[2]. When located in one city, it is often possible to order food from other cities. Hence, many restaurants appear more than ones in the dataset. When deleting duplicates, and only considering restaurants with a unique restaurant ID, the number of restaurants on Just Eat decreases to 2,193. On the frontpage of Just Eat's homepage it says: "*Gå på opdagelse i over 2.400 restauranter*", which translates to "*Explore more than 2,400 restaurants*" [Just Eat, 2020]. This information doesn't exactly match with the unique restaurant ID's extracted from the homepage. However, there could be few errors in the scraping process, even though it was conducted with great care and a lot of consideration. Another reason might be, that not all cities are included on the list provided by Just Eat. There definitely are more than 472 cities in Denmark.

When briefly investigating the Just Eat dataset, it is seen that some restaurant names are identical. In addition, a few restaurants even have the same address. Hence, a restaurant might appear twice in data, but with two different ID's. One reason is, that restaurant owners can add their restaurant multiple times, for instance by adding "Take away only" to the restaurant name. No further action has been taken to rectify this. Thus, a unique restaurant ID is considered as a unique restaurant.

To complete the data from Just Eat, information on the restaurant menus was gathered on December 13th, 2019, 10:29-10:56am. The Just Eat dataset will then include information

[1] An example of a city specific URL is: https://www.just-eat.dk/takeaway/aabenraa/?page=1
[2] Rating is displayed with a picture on the homepage why it has been excluded from the data.

on the price of a pizza Margherita, if this has been registered on the restaurant menu. In total, 680 of the 2,193 restaurant menus includes the word "Margherita". In some menus the word appears more than once, as a Margherita can be found in many different forms. This results in a few problems, as the prices might not 100 pct. reflect the same pizza. To try to be as consistent as possible, only the first Margherita price that appears on the web menu has been included in the dataset. However, on some menus, the first price reflects a special kind of price offer or pizza size, e.g. lunch offers or family sized pizza. In order to get prices which are as comparable as possible, prices, below 40 and above 70 DKK have been checked manually and corrected if necessary. The rest of the prices are assumed correct.

3.2 Geographical Data

As this paper attempts to investigate national pizza differences, geographical data is of great importance. The SAS software only provides a few different maps. Especially when looking at geographical regions instead of coordinate maps, SAS falls short. Luckily, it is easy to import any kind of shapefile (.SHP) map into SAS.

The Danish Agency for Data Supply and Efficiency distributes maps and geodata on the internet. Users are offered access to download free geographical data via a self-service map supply [Agency for Data Supply and Efficiency, 2020]. The dataset downloaded for this paper is the Danish Administrative Geographical Division (DAGI) in a 1:2 million scale. It is a standardized reference data, describing the country's administrative geographical division, and is updated each week. The data contains borders on the approx. 2200 parishes, 98 municipalities, 5 regions, 22 judicial districts, 12 police districts, 92 constituencies and approx. 1100 zip codes in Denmark [Agency for Data Supply and Efficiency, 2020]. The focus in this paper is the 98 municipalities. However, 99 groups are found in data, since Ertholmene, a small archipelago situated northeast of Bornholm, is included. As Ertholmene is not included in the municipal division, it will be excluded.

In order to be able to plot the Just Eat restaurants on a map in SAS, addresses in the Just Eat dataset need to be reviewed. Using the Danish Addresses Web API (DAWA) and some sophisticated SAS code, it is possible to get a lot of geographical information on each of the restaurants [SAS Community Nordic – Michael Larsen, 2014]. DAWA exhibits data and functionality regarding addresses in Denmark. The data provided by DAWA contains everything from administrative geographical information as well as land matrices on all Danish addresses [Danmarks Adressers Web API, 2020]. The administrative information provided by DAWA are perfect, as they make it easy to combine with the DAGI dataset.

The SAS program using the address API from DAWA tries to validate the inputted address'. Hence, an address needs to be specified exactly as in the DAWA database. As some restaurant address' are found to lack information (e.g. a letter indicator after a house number) and be misspelled, some restaurants can't be looked up using the address Web API. In total, 745 restaurants lack geographical information, and can't be plotted with exact coordinates. Fortunately, the restaurant information scraped from Just Eat include the zip code of the restaurant. A simple merge with a list of zip codes and it's belonging municipalities, makes it possible to include all 2,193 restaurants when the data are investigated on municipality level instead of restaurant level. The list of zip codes, updated on June 20th 2019, is downloaded from the PostNord homepage [PostNord, 2019]. PostNord is the leading provider of logistics solutions within the Nordic region.

When merging the restaurant zip codes with the list of zip codes including municipality information, some restaurants will be duplicated. This is due to some zip codes being in more than one municipality. In this dataset, 48 of the 2,193 restaurants are duplicated, and hence, checked manually, to make sure, the municipality information is as precise as possible. The final data shows, that seven of the 98 municipalities don't contain any restaurants. The municipalities without any Just Eat restaurants are Billund, Fanø, Lemvig, Læsø, Morsø, Vesthimmerlands and Ærø. In Table 1: , the four municipalities with most restaurants are displayed.

Table 1: Restaurant Frequency for Municipalities with more than 100 Restaurants

Municipality	**Restaurants**
København	356
Frederiksberg	265
Aarhus	221
Tårnby	130

3.2 Municipality Data

In order to try to explain why the price of pizza varies across the country, background data at municipal level is retrieved from StatBank Denmark. Statistics Denmark has established and maintains the StatBank as part of the dissemination of public statistics. The data are open and may freely be reproduced [Statistics Denmark, 2020]. However, as there are a lot of restrictions regarding data distribution, it is not possible to find all kind of data on StatBank Denmark. As an example, there are no tables on zip code level available for the

ordinary Dane. To get as much information as possible, data on municipality level has been selected.

In total, six different tables from the StatBank have been used to create a unified background data for all 98 municipalities. The tables are all with data from 2019 to make it as relevant as possible. They include FOLK1A and FOLK1C which provide information on the number of citizens, the gender, marital status as well as origin. The table GALDER provides an average age on all citizens, and in HFUDD10 the highest accomplished educational level can be found. In INDKP105 the average income is available. Finally, STRAF11 contains detailed information on all criminal actions reported to the authorities. The variables from all StatBank tables, including pizza prices divided on the 98 municipalities are shown in Table 2: .

Table 2: Average Number of Municipality Background Variables

Variable	N	Mean	Std Dev	Min	Max
Pizza Price	81	53,3518	53,2852	35,0000	75,0000
Total Number of Citizens	98	59.463	73.939	1.797	633.021
Prop. Men	98	0,4986	0,0079	0,4725	0,5131
Avg. Age	98	43,3939	2,9153	36,0000	53,6000
Prop. Divorced	98	0,0970	0,0127	0,0758	0,1250
Prop. Unmarried	98	0,4447	0,0466	0,3328	0,6527
Prop. Immigrants	98	0,0923	0,0384	0,0503	0,2473
Prop. Descendants	98	0,0285	0,0266	0,0045	0,1645
Prop. Primary School	98	0,2788	0,0463	0,1436	0,3734
Prop. Long HE	98	0,0744	0,0543	0,0214	0,2717
Avg. Income	98	325.276	54.926	259.870	594.637
Prop. Violations	98	0,0093	0,0204	0,0002	0,1919
Prop. Arson	98	0,0042	0,0070	0,0000	0,0526
Prop. Shoplifting	98	0,0293	0,0208	0,0000	0,1103
Prop. Moped Theft	98	0,0028	0,0034	0,0000	0,0233
Prop. Bicycle Theft	98	0,0910	0,0580	0,0000	0,2918
Prop. Vandalism	98	0,0582	0,0305	0,0000	0,2558

Note: The proportions are all calculated from the total number of citizens in the municipalities. The proportion of violations are calculated from the total number of violations in the entire country.
HE = Higher Education

4 The SAS® Platform

Data preparation and analyses in this paper was conducted using the SAS® Platform, consisting of SAS® 9.4 and SAS® Viya®. Both contribute with software components that integrate into a unified SAS Platform. The data collection and structuring were primarily done using SAS 9.4, while the data plotting and analyses were done with the new SAS Viya.

4.1 Essential Procedures

Even though SAS 9.4 programs can run unmodified and faster from within SAS Viya, the initial data collection was done using a desktop client, SAS® Enterprise Guide. In this way, the data generated is saved locally on the computer. Hence, the final modified data, need to be uploaded on the SAS Viya CAS server, when ready for analyses.

The MAPIMPORT procedure comes in handy when needing to expand the available map data that are already available with SAS/GRAPH [SAS Institute Inc., 2017]. The code for importing a SHP file is as follows:

```
proc mapimport
     datafile="…\MUNICIPALITIES_DK.shp" out=Municipality;
     id Municipality_name;
     select Municipality_code Region_name Region_code;
run;
```

If an identification variable is already given in the SHP file, this can be specified using an ID statement. If the dataset contains redundant variables, the SELECT statement will include only the specified variables. Further, variables with shape information for the polygons that compose the map and the identification information are included.

When the polygonal maps are in place, it is possible to add response data, that contain a matching ID. In this case, the response data of interest is gathered from the Just Eat homepage. When conducting web scraping in SAS, the HTTP procedure is needed:

```
filename source &fileref;
proc http
     url="https://www.just-eat.dk/takeaway/&city/?page=&pagenum"
     out=source;
run;
```

The URL statements identifies the URL path for the homepage of interest. In the example above, the source files are stored on the local computer as text files, which can then be imported into SAS. In this process, some understanding of html programming is an

advantage. Macro coding is essential, as the web scraping process is repeated for multiple URLs.

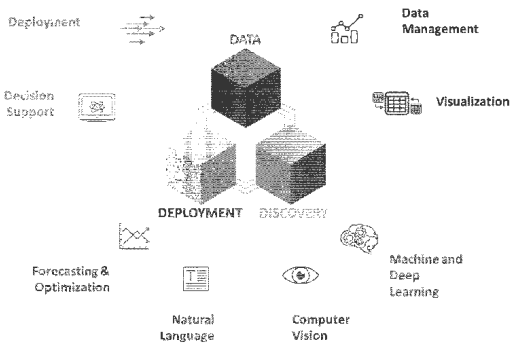
As the HTTP procedure is not supported on the CAS server, which is preferred by SAS Viya, this is the main reason, why the initial data preparation was done using SAS 9.4. However, some settings can be modified, in order to re-enable the procedure [SAS Institute Inc., 2019A].

4.2 The Analytics Life Cycle

SAS® Viya® is the newest addition to the SAS Platform. It has a lot of advantages: By multi-threading the data step, meaning that data rows are processed in parallel batch SAS sessions and rejoined after all sessions have completed, the speed of running SAS code increases tremendously. Additionally, the SAS® Cloud Analytic Services (CAS), a server that provides security, allows users to share data between sessions and provides fault tolerance, further improve the run-time [Jerry Pendergrass, SAS Institute, 2017]. SAS Viya is open, in the sense that it integrates with open source, so that external interfaces, like R or Python, can take control of SAS Cloud Analytic Services from Jupyter Notebook [Phil Weiss, SAS Users Blog, 2017].

Accessing the SAS Viya interface reveals additional value for the SAS user, as SAS Viya integrates with the SAS analytics life cycle. The analytics life cycle, depicted in Figure 2, represents the journey from data to intelligence. It contains everything from data, discovery to deployment.

Figure 2: The Analytics Life Cycle



The interface makes it easy for the user, as one can for instance select "Build Models" from a functional menu, cf. Figure A1 in Appendix, without having to know that they are leveraging the SAS Visual Data Mining and Machine Learning product [SAS Institute Inc., 2019B].

5 Results

Combining the data on Just Eat restaurants with the geographical municipalities' division, it is possible to get a better overview of the Danish pizza consumption.

5.1 Just Eat Pizza Geography

In total, 1420 of the 2,193 restaurants from Just Eat have a unique set of geographic coordinates, which can be plotted using SAS® Visual Analytics on SAS Viya. Only a brief look at Figure 3 shows that most restaurants are located near the big cities. Further, the price, indicated by the color of the stars (blue being more expensive than red, and gray indicating that no price was registered for the given restaurant), shows that the most expensive pizzas are found near the big cities. Notice, that the price of a Margherita pizza was only found for 659 of the restaurants.

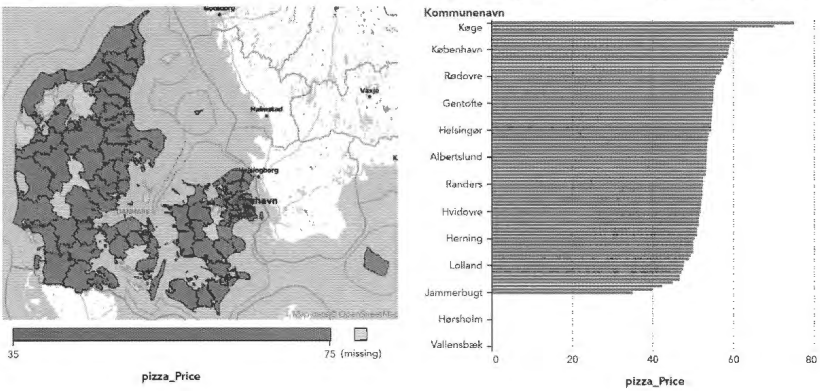
Figure 3: Coordinates Map of Restaurants

Note: This geographical plot in SAS Visual Analytics on SAS Viya uses a set of style options. Hence, the markers have been changed to stars instead of the default circle, and the map background has been changed to the World Dark Gray Base canvas instead of the

OpenSteetMap. Finally, the price of Margheritas has been added as a color scale.

When changing the geo coordinates to a geographical region, the restaurants, and hence the pizza price can be investigated on a municipality level. In order to get the region plotted in SAS Visual Analytics, it is necessary to create a new geography data item. Here, the dataset from DAGI brought into SAS by the MAPIMPORT procedure is needed, as it contains latitude and longitude coordinates for all points encircling the various municipalities. In Figure 4, the municipalities are displayed as geographic regions with colors reflecting the price of Margherita pizzas. Additionally, the average pizza prices are displayed in a bar chart for all 81 municipalities[3].

Figure 4: Region Map and Bar Chart of Average Margherita Prices by Municipalities



In SAS Viya, it is easy to zoom in and out in order to get a better look on all the different municipalities. What is clear from Figure 4, is that a lot of the restaurants have Margherita pizzas with an average just around 50 DKK. However, a few municipalities stand out. In the northern part of Jutland, the pizza prices are just around 35-40 DKK, which is generally lower than in the rest of the country. In the two municipalities, Køge and Nordfyns, the prices are a lot higher, 75 and 70 DKK respectively. Hence, the difference in consumer pricing (if comparing the Margherita prices in Denmark to the Big Mac Index) is to some extent present in Denmark. Especially, the low prices in the northern part of Jutland makes sense, when examining restaurant sales from 2018 provided by Statistics Denmark. Both

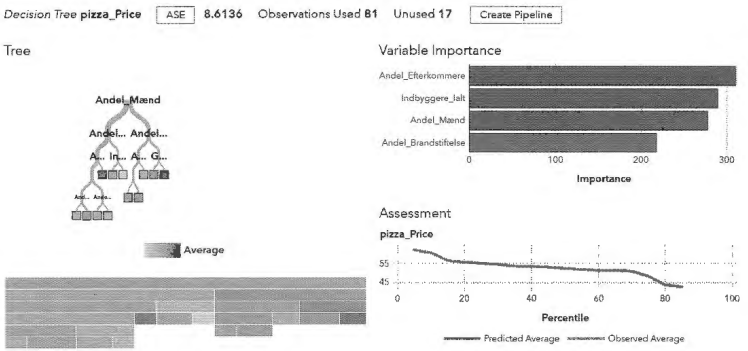
[3] The following 17 municipalities don't have any restaurants with specified pizza prices: Ærø, Billund, Dragør, Faaborg-Midtfyn, Fanø, Fredericia, Glostrup, Hørsholm, Ikast-Brande, Læsø, Lejre, Lemvig, Morsø, Ringsted, Samsø, Vallensbæk and Vesthimmerlands.

total sales as well as the number of employees in the restaurant industry are lowest for the region of northern Jutland [Statistics Denmark, 2019].

5.2 Municipality Characteristics

The reason for the varying pizza prices across the country might not be a result of the difference in PPP. There might be a lot of reasons why the prices differ, such as supply and demand within the region. The 16 variables from Table 2 in Section 3.2, including the number of Just Eat restaurants within each municipality, are used to investigate whether any municipality specific characteristics can explain the Margherita prices. Using SAS® Visual Statistics in SAS Viya, it is possible to create a decision tree, as shown in Figure 5, which can show the importance of the variables.

Figure 5: Results from Decision Tree Model



What is clear from the output in Figure 5 are the four most important variables when explaining pizza prices. According to the model, the proportion of immigrant descendants is most important, followed by the total number of citizens, the proportion of men and the proportion of arson of the total number of violations in a given municipality[4]. The root node is the proportion of men and is split into two groups; municipalities with a proportion of men larger than 49.7 pct. and municipalities with a proportion of men between 47.4 and 49.7 pct. In Table Table 3, the two leaf nodes with the highest and lowest price are given.

[4] Arson is the criminal act of deliberately setting fire to property

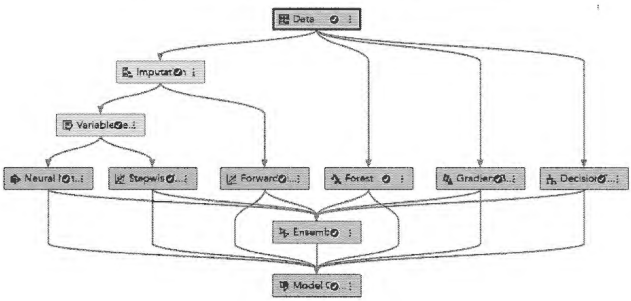
Table 3: Decision Tree Nodes with Highest and Lowest Pizza Price

Highest Avg. Price	**62.1873**	**Lowest Avg. Price**	**42.8000**
Std. Dev.	6.3278	Std. Dev.	5.8243
Number of Municipalities	6	Number of Municipalities	5
Prop. Arson	≤ 0.002	Prop. Divorced	0.076 - 0.086
Prop. Men	0.474 - 0.497	Prop. Men	≥ 0.497
Avg. Income	259,870 - 330,926	Number of Citizens	12,505 - 43,435.2

The nodes with the highest average price, 62.19 DKK, contains 6 municipalities. A split happens three times. First the proportion of arson of the total number of violations within a given municipality is below 0.002. The proportion of men is between 0.474 and 0.497 and the average income is between 259,870 and 330,926 DKK in a year. Hence, a low proportion of arson, few men and a relatively low income result in high Margherita prices. On the contrary, a low average pizza price of 42.80 DKK is characterized by a low proportion of divorced citizens, a high proportion of men and a relatively low number of citizens.

The gradient boosting model can be created just like the decision tree from the "Explore and Visualize" menu in SAS Viya. But instead of using the SAS® Visual Data Mining and Machine Learning, "Build Models" are selected in the SAS Viya menu, which direct the user to SAS® Model Studio.

Figure 6: SAS® Model Studio Pipeline



In SAS Model Studio, it is possible to create predictive and classification models using pipelines, which are very similar to SAS® Enterprise Miner projects. In Figure 6 a pipeline template has been used, to create six different machine learning models, as well as an ensemble model. The models are easily created from an input dataset, and compared in the

Model Comparison node. According to the model comparison, the gradient boosting model is the champion model. When investigating the gradient boosting model node, a list of all input variables are listed by importance. In Table 4 the six most important variables are displayed. According to this model, the most important variable is the frequency of Just Eat restaurants from the analysis. The second most important is the number of descendants, and maybe a bit surprising, the proportion of two violation types affect the average price of a pizza Margherita.

Table 4: Variable Importance Based on Gradient Boosting Model

Variable	Training Importance	Importance Std. Dev.	Relative Importance
Number of Restaurants	18.5881	33.8333	1
Prop. Descendants	17.0882	51.8182	0.9193
Prop. Vandalism	14.7882	49.3732	0.7956
Prop. Bicycle Theft	11.2469	71.4004	0.6051
Prop. Men	10.3079	35.8661	0.5545
Prop. Divorced	9.8372	39.7822	0.5292

Note: The training dataset was assigned 100 pct. of the data in the partitioning.

6 Conclusion

Food has a large effect on the world. Something as simple as a burger can help us understand differences in economic productivity and standards of living between countries. In this article, the average price of pizza Margheritas across the 98 municipalities in Denmark were investigated. It turns out, that there is a difference in prices across the nation. Generally the price lies around 50 DKK for most municipalities. By creating a geographic plot and comparing the prices, it is clear, that pizzas are a lot lower in northern Jutland.

When examining which municipality characteristics affect the price most, the number of restaurants within a specific municipality is important. Additionally, the proportion of immigrant descendants and proportion of men within a municipality are essential. However, this article is only based on Margherita prices from 659 restaurants across Denmark. Hence, the results are not very robust, but provides an indication of how it looks. In order to conclude something with more confidence, a broader analysis need to be conducted.

Nevertheless, it is always important to remember, that food is much more than just a means to quench your hunger.

References

Agency for Data Supply and Efficiency (2020). *The Danish Map Supply*. Retrieved 6 January 2020, from https://kortforsyningen.dk

Danmarks Adressers Web API (2020). *Danmarks adresser*. Retrieved 6 January 2020, from https://dawa.aws.dk/

Jerry Pendergrass, SAS Institute (2017). *The Architecture of the SAS® Cloud Analytic Services in SAS® Viya™*. Retrieved 6 January 2020, from https://support.sas.com/resources/papers/proceedings17/SAS0309-2017.pdf

Just Eat (2020). *Just Eat*. Retrieved 6 January 2020, from https://www.just-eat.dk

Phil Weiss, SAS Users Blog (2017). *Top 12 Advantages of SAS Viya*. Retrieved 6 January 2020, from https://blogs.sas.com/content/sgf/2017/10/20/top-12-advantages-of-sas-viya/

PostNord (2019). *Postnummerkort og postnummerfiler*. Retrieved 6 January 2020, from https://www.postnord.dk/kundeservice/kundeservice-erhverv/om-postnumre/postnummerkort-postnummerfiler

SAS Community Nordic – Michael Larsen (2014). *Juletip #19 Adresse validering ved hjælp af filename URL*. Retrieved 6 January 2020, from https://communities.sas.com/t5/Nordic-Events-and-Presentations/Juletip-19-Adresse-validering-ved-hj%C3%A6lp-af-filename-URL/ba-p/223472

SAS Institute Inc. (2017). *MAPIMPORT Procedure*. Retrieved 6 January 2020, from https://go.documentation.sas.com/?docsetId=grmapref&docsetTarget=p1rvpewhocha4yn1gipf920vchn3.htm&docsetVersion=9.4&locale=en

SAS Institute Inc. (2019A). *HTTP Procedure*. Retrieved 6 January 2020, from https://go.documentation.sas.com/?docsetId=proc&docsetTarget=n0bdg5vmrpyi7jn1pbgbje2atoov.htm&docsetVersion=9.4&locale=en

SAS Institute Inc. (2019B). *SAS® 9.4 and SAS® Viya® Functional Comparison*. Retrieved 6 January 2020, from https://support.sas.com/resources/papers/sas-94-sas-viya-functional-comparison.pdf

SAS Software YouTube Channel (2017). *Decision Trees, Boosting Trees, and Random Forests: A Side-by-Side Comparison*. Retrieved 6 January 2020, from https://www.youtube.com/watch?v=gehNcYRXs4M

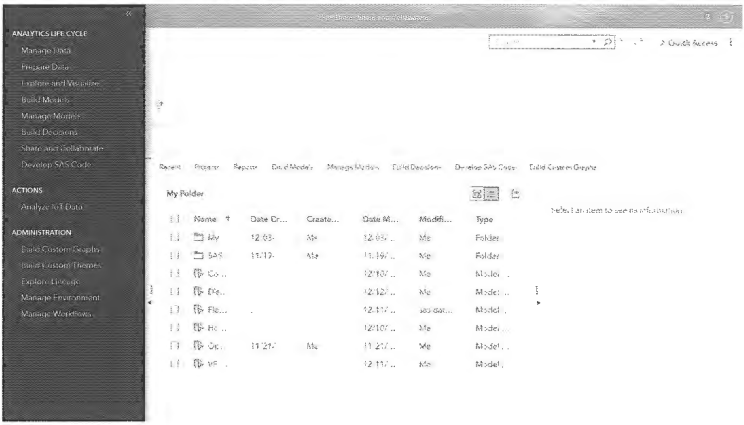
Statistics Denmark (2019). *Restaurationsbranchen vokser fortsat i hele landet*. Retrieved 6 January 2020, from https://www.dst.dk/da/Statistik/bagtal/2018/2018-02-19-restaurationsbranchen-vokser-i-hele-landet

Statistics Denmark (2020). *Statistikbanktabeller på kommuneniveau*. Retrieved 6 January 2020, from https://www.dst.dk/da/Statistik/kommunekort/statistikbanktabeller-paa-kommuneniveau

The Economist (2019). *The Big Mac index*. Retrieved 6 January 2020, from https://www.economist.com/news/2019/07/10/the-big-mac-index

Appendix

Figure A1: The Menu in SAS® Viya®



Innovation og kreativitets effekt på økonomisk performance

Analyse af robusthed af resultater eller om at sælge elastik i metermål?

Mogens Dilling-Hansen
Department of Economics and Business Economics
Aarhus Universitet
Mail: dilling@econ.au.dk

Resumé

Små kreative virksomheder er udset til at drive økonomisk udvikling de kommende år. Nordic-Buzz projektet viser, at der er stadig er et stykke vej til den situation. Ikke desto mindre viser resultaterne i denne analyse, at det faktisk er de kreative virksomheder, der klarer sig relativt bedst.
Det grundlæggende problem ved analyser af denne slags er, at omsætningstal er så begrænsede at de statistiske analyser ikke giver klare resultater, og specielt peges der på, at et væsentligt problem er den måde, som virksomhederne kategoriseres på.

1 Indledning

"*Det kræver ærlighed at sælge elastik i metermål*" er et ofte citeret ordsprog af Storm P. Ordsprog er der mange af, lige som der i den akademiske verden er gjort utallige forsøg på at forklare hvordan menneskelig kreativitet har en positiv effekt på økonomisk udvikling, se fx Kaufman et al (2010).

Umiddelbart var det ovenstående ordsprog, der dukkede op i hukommelsen, efter at have læst den indledende artikel til bogen "*Creativity & Innovation. Theory, Research and Practice*" af Plucker (2017), hvor kreativitet har udviklet sig fra at være en hippie-ting til at være en vigtig determinant for fortsat stigning i mængden af innovation. Samme tema er fremherskende i vurderingen af de kreative erhvervs ernorme potentiale for at skabe fremtidig økonomisk vækst, se fx. Dansk Erhverv (2019): Kreativitet og de kreative erhverv vil blomstre de kommende år og et sandt væksteventyr venter for hele økonomien!

Analyser af SMV'er inden for de kreative erhverv viser et noget andet billede, og det gælder fra helt basal mangel på økonomisk vækst til en manglende økonomisk fokus for virksomheden. Formålet med denne analyse er at påvise, hvorledes analyser af virksomheder er kraftigt påvirket

af kvaliteten af de anvendte data, og det er særligt aktuelt, når der anvendes begreber som kreativitet og innovationer; ofte er definitionen ret utvetydig, medens måling af begreberne er lang vanskeligere.

Opbygningen af analysen er som følger. Kapitel 2 indeholder en redegørelse af relevante definitioner af begreber med en diskussion af måleproblemer, og der fortsættes i kapitel 3 med en præsentation af en model for sammenhæng mellem kreativitet, innovation og økonomisk performance. Kapitel 4 er en præsentation af det eksperimentelle analysedesign med tre kohorter af små kreative virksomheder, hvorefter analyserne præsenteres i kapitel 5. Kapitel 6 runder af med en diskussion af de fundne resultater.

2 Begrebsafklaring mht. kreativitet og innovation

Formålet med denne analyse er at forklare, hvorfor nogle små kreative virksomheder klarer sig bedre end andre, og derfor er det nødvendigt at definere begrebet *økonomisk performance*. En stor del af denne virksomhedsgruppe er relativt nystartede virksomheder, så derfor anvendes meget simple mål for performance i form af data baseret på virksomhedens omsætning og beskæftigelse. I alle tilfælde er begreberne relativt klare og usikkerheden for disse mål er hovedsalig baseret på det forhold, at oplysningerne i denne analyse er baseret på selvrapporterede informationer.

Kreativitet er typisk er personlig egenskab, der er god til at få nye ideer og realisere dem på en fantasifuld eller kunstnerisk måde (denne version er taget fra "Den danske Ordbog"). I mange sammenhænge er ordet brugt som tillægsord for at beskrive genstandens (virksomhed, klasse, bogføring…) for at beskrive den lidt alternative måde, et problem løses på. I den økonomiske litteratur er definitionerne også mange, spændende fra de helt klassiske definitioner hvor kreativitet fører til unik poesi, kunst og musik (Platon, Sokrates, Aristoteles mfl.) til noget originalt, brugbart og overraskende for individ/samfund, se oversigt i Dow (2017). Der er også meget stor forskel på typen af kreativitet, og her tænkes specielt på sondringen *divergent / convergent creativity*, hvor den konvergente kreativitet er specielt interessant, fordi den bygger på traditionelle strukturerede metoder til at komme frem til eet resultat (lineær, systematisk metode), se Cortes et.al. (2019).

I alle tilfælde er definitionen af kreativitet relativ let at forstå, men meget svær at kvantificere, og i ovenstående to ekstremer er det ret klart, at kun de nyere definitioner er kvantificerbare; problemet er blot, at definitionerne er meget tæt på definitionen af innovation, se fx diskussionen i Dino (2017), hvor kreativitet defineres som skabelsesprocessen og innovation som implementeringsdelen af en innovativ aktivitet. Under alle omstændigheder er konklusionen mht. måleligheed, at den faktiske eksistens af kreativitet er mulig at måle, men omfanget af samme er kun mulig via indikatorer (ordinal information).

Innovation er modsat kreativitet knyttet til et produkt eller en ydelse, således innovation dækker over både udvikling og praktisk realisering af en ny ide. I tråd med Eurostat/OECD (OECD/Eurostat (2019)), definerer Danmarks Statistik (2019) innovation i private virksomheder som "*introduktionen af et nyt eller væsentligt forbedret produkt (vare eller tjenesteydelser), proces, organisationsform eller markedsføringsmetode*". Definitionen af innovation er naturligt rettet mod private virksomheders innovation, og der er en række uddybende af beskrivelser af innovationstyper og –former, se fx Tidd et.al. (2005); fælles for dem alle er, at de betoner vigtigheden af at kun teknologiske nyskabelser, der får en kommercielt udbredelse, kan anses for at være innovation.

De måleproblemer, der er identificeret ved opgørelse af kreativitet, er også at finde ved opgørelse af innovative aktiviteter; men til gengæld er der udover opgørelse af innovation ved indikatorer også mulighed for at måle udgifterne og beskæftigelse ved at skabe innovation. Opgørelse af udgifter til udvikling og personale i udviklingsafdelingen (også kaldet "innovation efforts" beskrevet i OECD (2015)) er en meget præcis indikator for omfanget af innovation, men metoden lider af to fundamentale problemer. For det første er det den formelle del af innovation, R&D (Research & Development), kun en ægte delmængde af den samlede innovationsaktivitet i en virksomhed, og for det andet er det primært større virksomheder, der har en decideret R&D-afdeling, hvor udgifter kan opgøres usikkerhed.

At der er måleproblemer ved opgørelse af kreative og innovative aktiviteter er ikke overraskende og ej heller invaliderende for analyser af innovation, da langt de fleste opgørelsesmetoder er ordinale af natur; det er muligt at kvantificeringen ikke er helt korrekt, men retningen af effekterne er troværdige.

Mere problematisk er det, at begreberne kreativitet (som normalt er et individuel karakteristika) og innovation (som normalt opgøres på virksomhedsniveau) ikke er uafhængige af hinanden og at der grundlæggende er tvivl om den indbyrdes sammenhæng. På den ene side er der en *innovation management* litteratur, der naturligt anerkender kreativitet som en del af innovationsprocessen, men som også forudsætter kreativitet som en *core attribute* af den innovative person, se fx Tidd et.al. (2005) – med andre ord er kreativitet en uafhængig størrelse, der kan opgøres uafhængigt af innovationsomfanget. På den anden side er der i nyere innovationslitteratur fokus på innovationsprocessen, hvor kreativitet er en del af skabelsesprocessen, se Dino (2017) – i dette tilfælde er opgørelse af kreativitet og innovation ikke uafhængige størrelser.

I det efterfølgende afsnit præsenteres et bidrag, hvor effekten på virksomhedens performance (økonomisk omsætning) af den samlede indsats af kreativitet og innovation opstilles og analyseres. I kapitel 4 testes modellen på små skandinaviske virksomheder for at vurdere modellens implicitte antagelse af uafhængighed mellem kreativitet og innovation.

3 En model for små kreative virksomheders udvikling

I Chaston (2008) undersøges 107 små kreative virksomheder og data er survey-baserede. Formålet med analysen er at finde ud af, hvorvidt formelle uddannelsesprogrammer for de små virksomheder faktisk skaber økonomiske vækst.

Diskussionen af sammenhængen mellem succes for de små kreative virksomheder og succes på et makroøkonomisk niveau er ganske interessant; men det egentlige bidrag i den teoretiske diskussion er, at det er muligt at opstille tre hypoteser, der relaterer økonomisk performance/succes med kreativitet og innovation.

Den første hypotese er baseret på en række observationer af kreative virksomheder, der grundlæggende er drevet af kreativiteten mere end af at tjene penge. Det er ikke nødvendigvis sådan at virksomheden er enten eller, men derimod opfattes de to begreber (kreativitet-økonomi) som to ekstremer på en ordinal 6-punkts skala, og det fører til følgende H_0-hypotese

H1: Små kreative virksomheder er drevet af at skabe finansiel succes.

Den anden hypotese er baseret på det forhold, at mange små virksomheder mere er entreprenører end leder af en virksomhed, der har profit og vækst for øje. Entrepreneurielle virksomheder har

ifølge Chaston (2008) stor fokus på at markere sig som en succesfuld iværksætter, som tilbøjelig til at tage for store risici og specielt for at være (for) fokuseret i innovation. På tilsvarende måde som for den kreative hypotese opstilles følgende H_0-hypotese om at iværksætter-dimensionen ikke har betydning

H2: Små kreative virksomheder har ingen fokus på entrepreneurielle forhold.

Den sidste hypotese, der opstilles, er baseret på en diskussion af, hvorvidt den entrepreneurielle og kreative dimension faktisk påvirker hinanden, og det fører til den tredje hypotese

H3: Små kreative virksomheders kreative og entrepreneurielle dimensioner er uafhængige

Ved at gruppere virksomhederne i de to dimensioner (kreative/artistiske og entrepreneurielle dimension) i binære variable skabes der fire virksomhedstyper:

Tabel 1 Performance baseret på hhv. kreativ og entrepreneuriel orientering

Virksomhedstype	*Konventionel*	*Entrepreneuriel*
Finansiel	Performance ++	Performance ?
Kreativ	Performance ?	Performance --

Kilde: Baseret på Chaston (2008)

Tabel 2 Resultater baseret på model i tabel 1

Virksomhedstype	*Konventionel*	*Entrepreneuriel*
Finansiel	Andel 27% Performance 3.98	Andel 9% Performance 4.09
Kreativ	Andel 46% Performance 2.71	Andel 18% Performance 2.23

Kilde: Baseret på Chaston (2008)

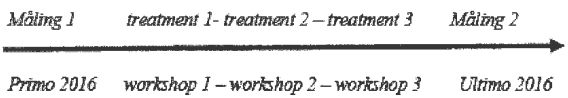
Resultaterne i tabel 2 viser dels den relative fordeling af de 107 virksomheder og dels den gennemsnitlige performance for hver af de fire grupper – bemærk der måles på en skala fra 1 til 6 og at der grundet den relativt lille og skæve stikprøve ikke er signifikant forskel på tallene; men hvis der alene testes på forskelle mellem finansielle og kreative virksomheder, så er forskellen signifikant; baseret på denne undersøgelse kan H1 og H2 ikke afvises, men H1afvises: Det er kreativiteten, der forstyrrer den nødvendige fokus på finansiel/økonomisk succes.

4 NordicBuzz 2016-18 – et eksperimentelt analysedesign

I perioden 2016 til 2018 blev en række små danske og svenske kreative virksomheder analyseret med henblik på at vurdere, om det er muligt at øge den finansielle/økonomiske fokus for denne type ved at uddanne virksomhederne. Der er med andre ord tale om en helt klassisk problemstilling, hvor effekten af et indgreb skal vurderes (*treatment effekt*), og for at kontrollere for heterogeniteten blandt de deltagende virksomheder er der anvendt et eksperimentelt analysedesign med måling før og efter tre workshops, som de tre kohorter deltog i det første analyseår. De tre workshops, der blev afviklet i hhv. 2016, 17 og 18, var designet således de matchede efterspørgslen efter økonomisk indsigt (eller mangel på samme), og deltagernes evaluering af alle workshops var meget positiv.

Simpel deskriptiv statistik for de tre kohorter kan ses i tabel 3, og der tegnes for alle tre år et meget typisk billede af de små kreative virksomheder: Den kreative dimension har meget stor indflydelse på forretningsmodellen og der er på den korte bane meget lidt fokus på omsætning, vækst, skalering af virksomhed, ledelse og generelle økonomiske udfordringer, der gælder for alle relativt nystartede virksomheder.

Figur 1 Eksperimentelt design for NordicBuzz-analyserne illustreret for 2016-kohorten



Kilde: Dilling-Hansen & Andersen (2018)

For at måle udviklingen på kort sigt er der i Dilling-Hansen & Andersen (2018) opstillet en GAP-model til måling af udviklingen. Ideen er, at virksomhederne selv vurderer deres aktuelle og ønskede status på seks skalaer, der vedrører virksomhedens driftsøkonomiske aspekter: Jo større forskel mellem den ønskede position og den faktiske position, jo større gap; jo større gap de konstateres, jo mere er der at rette op på. Målingen af det "driftsøkonomiske gap" foregår som beskrevet i figur 1 både før og efter de tre workshops, og det er dermed udviklingen i gap'et, der angiver at det går i den rigtige retning. Som det ses af tabel 3, så viser gap-analyserne, at denne del af projektet er vurderet som meget vellykket.

Tabel 3 NordicBuzz, beskrivende statistik for 2016-18 kohorter

Kohorte	2016	2017	2018
Antal virksomheder	26	21	29
Andel kvindelige ejere	65,4%	76,2%	75,9%
Ejers alder, år	36,2	38,3	35,5
Ejers civilstand, andel enlige	30,8%	33,3%	34,5%
Virksomhedens alder, antal år	1,8	3,2	2,1
Virksomhedsdata EX-ANTE treatment			
Virksomhedens omsætning, år t-3, 1000 kr	32	21	36
Virksomhedens omsætning, 1000 kr	167	194	390
Virksomhedens omsætning, år t+3, 1000 kr	1.586	1.791	6.095
Virksomhedens beskæftigelse	0,6	0,6	1,0
Virksomhedens gap	1,91	2,02	1,52
Virksomhedsdata EX-POST treatment			
Virksomhedens omsætning, 1000 kr	368	268	399
Virksomhedens beskæftigelse	0,7	0,8	1,1
Virksomhedens gap	1,59	1,62	1,30

Noter: NordicBuzz-data, se Dilling-Hansen & Andersen (2018)

For at vurdere kreativitet, innovation og entrepreneuriel adfærd har virksomhederne har vurderet graden af enighed på en 7-punkts Likert-skala for følgende spørgsmål

K.1 Jeg anser mig selv for at være en kreativ person

K.2 Jeg anser min virksomhed/företag samt dets produkter for værende kreativ(e)

K.3 Jeg anser min virksomhed/företag for værende innovativ (nye produkter, produktionsformer, salgsformer)

E.1 Økonomisk succes er langt vigtigere end personlig udvikling og kreativitet

E.2 Økonomisk succes er langt vigtigere end kreativ udvikling af min virksomhed

E.3 Økonomisk succes er langt vigtigere end at have en god balance mellem privat liv og arbejde

E.4 Økonomisk succes er langt vigtigere end at afspejle min kunstneriske baggrund

E.5 Jeg er villig til at gå på kompromis med mine produkter/idéer/kreativitet for at kunne skabe økonomisk succes for min virksomhed

E.6 Det betyder meget for mig, at jeg som iværksætter skaber vækst og udvikling for samfundet

E.7 Det betyder meget for mig, at jeg skaber en bæredygtig virksomhed

De stillede spørgsmål er inspireret af Chaston (2008) og generelle overvejelser om måling af de tre dimensioner. Der er ikke data for 2016-kohorten, og svarene på de 10 spørgsmål er fordelt som forventet (svagt venstreskæve fordelinger) for virksomheder, der er erklærede kreative virksomheder. Faktoranalyse er brugt til at identificere latente strukturer (Malhotra et.al. (2012)), og

analyserne viste, at der ikke var uafhængighed mellem de kreative og de entrepreneurielle spørgsmål, jf. diskussion i slutningen af kapitel 3.

5 Analyse af performance, kreativitet og innovation

På baggrund af faktoranalyserne i kapitel 4 er spørgsmål E.1, E.2 og E.3 valgt som indikator for kreativitet og K.3 valgt som indikator for entrepreneurielt syn, og for begge indikatorer er midtpunktet valgt som afskæringspunkt for dannelse af den binære variabel.

Baseret på data for virksomhederne i 2017- og 2018-kohorten (spørgsmålene er ikke stillet på samme måde i 2016-stikprøven) er sammenhængen mellem kreativitet og innovation opstillet i tabel 4. Fordelingen af de 50 virksomheder er i ex-ante opgørelsen (til venstre) opgjort før en evt. treatment-effekt (tre workshops) og tabellen ex-post (til højre) er samme opgørelse opgjort efter evt. treatment effekt.

Tabel 4 Inddeling i virksomhedstyper baseret på ex-ante og ex-post vurdering

Ex-ante	*Konventionel*	*Innovativ*
Finansiel	3 (6%)	7 (14%)
Kreativ	14 (28%)	26 (52%)

Ex-ante	*Konventionel*	*Innovativ*
Finansiel	3 (6%)	7 (14%)
Kreativ	15 (30%)	25 (50%)

Kilde: Baseret på Chaston (2008) og NordicBuzz-data. I både ex-ante og ex-post analysen forkastes hypotesen om uafhængighed mellem inddelingskriterierne. I opgørelsen ex-ante og ex-post af de 50 virksomheder har 18 virksomheder skiftet kategori.

Tabel 4 indeholder ingen performance data; men ikke desto mindre er der en række markante resultater. For det første er de 50 virksomheder kreative virksomheder, og derfor er der kun 20% af de 50 virksomheder, der svarer at de agerer efter finansielle/økonomiske motiver! I Chaston (2008) er denne andel på 36%.

For det andet er der ikke uafhængighed mellem de to inddelingskriterier i begge tilfælde; men den tilsyneladende stabile fordeling dækker over, at 18 ud af de 50 virksomheder har skiftet holdning til, hvorvidt de anser virksomheden for at være overvejende kreativ hhv. innovativ.

Spørgsmålet er, om denne store variation i grundlæggende opfattelse af virksomhedens type, også har effekt på analysen af virksomhederne performance?

I stil med Chaston (2008) er der her brugt omsætningstal for virksomhederne, men der præsenteres to omsætningstal for at fange treatment-effekten på omsætningen: Jf. tabel 1 er der i løbet af

analyseåret sket en ret markant stigning i omsætningen – en stigning som virksomhedsdeltagerne selv har skabt.

Tabel 5 Virksomhedsperformance baseret på ex-ante og ex-post vurdering

Ex-ante	*Konventionel*	*Innovativ*	*Ex-ante*	*Konventionel*	*Innovativ*
Finansiel	40.800	102.333	***Finansiel***	8.133	287.907
	14.400	182.667		14.400	72.000
Kreativ	320.088	372.248	***Kreativ***	464.900	242.220
	427.200	242.649		408.000	260.653

Kilde: Baseret på Chaston (2008) og NordicBuzz-data. I både ex-ante og ex-post analysen forkastes hypotesen om uafhængighed mellem inddelingskriterierne. I opgørelsen ex-ante og ex-post af de 50 virksomheder har 18 virksomheder skiftet kategori.

Resultaterne i tabel 4 og 5 viser, at der er god grund til at være forsigtig med operationalisering af variable til inddeling af virksomheder i grupper, og specielt er tolkningen vanskelig af det store antal virksomheder, der faktisk skifter opfattelse af, hvilken virksomhedstype, den faktisk er: En venlig tolkning er, at virksomhederne faktisk skifter adfærd som følge af de workshops, som de har deltaget i (treatment-effekt); men det kan også udtrykke en reel usikkerhed i data, som kun reddes af at performance-analyserne ikke er signifikante.

Datamaterialet er begrænset til 50 virksomheder, og derfor er ingen 2-factor ANOVA analyser signifikante; men selv om forskellene – helt i tråd med Chaston (2008) – ikke er signifikante, så er forskellene ret markante – og i klar kontrast til Chaston (2008) - når de kreative erhverv sammenlignes med de virksomheder, der i overvejende grad styres af økonomiske motiver; de kreative virksomheder er kreative fordi de faktisk er gode til det!

6 Afrunding

"Det jævner sig…"; styrken ved anvendelse af statistiske analyser er netop, at mindre – måske tilfældige – afvigelser ikke får indflydelse på konklusionerne.

NordicBuzz projektet har vist, at det faktisk er en god ide at påvirke de små kreative virksomheder, så de bevæger sig i en mere holdbar situation med større fokus på økonomien.

Det fjerner heldigvis ikke de kreative virksomhedes fokus, som i grel modsætning til, hvad teorien forventer, faktisk er i stand til at skabe omsætning!

Referencer

Chaston, I. (2008), "Small creative industry firms: A development dilemma?", *Management Decisions*, 2008, vol. 46, no. 6: 819-831.

Cortes, R.A,, A.B. Weinberger, , R.J. Daker & A.E. Green (2019), "Re-examining prominent measures of divergent and convergent creativity", *Current Opinion in Behavioral Science*, 2019, No 27: 90-93.

Danmarks Statistik (2019), "*Innovation*", Danmarks Statistik, 2019. Kilde: https://www.dst.dk/da/Statistik/emner/uddannelse-og-viden/forskning-udvikling-og-innovation/innovation

Dansk Erhverv (2019), "Godt regeringsinitiativ for vækst i de kreative erhverv", online *https://www.danskerhverv.dk/presse-og-nyheder/nyheder/godt-regeringsinitiativ-for-vakst-i-de-kreative-erhverv/*

Dilling-Hansen, M. & O. L. Andersen (2018) "*Nordic Buzz 2016-18. Udvikling af kreative iværksætter virksomheder*", Lifestyle & DesignCluster, 2018. Online https://ldcluster.com/?s=nordic+buzz

Dino, R.N. (2017), *Connected but Different. Comparing and Contrasting Creativity, Innovation, and Entrepreneurship*, in Plucker (ed.), "Creativity and Innovation: Theory, research and practice" Prufrock Press Inc. 2017, ISBN 978-1-61821-595-6.

Dow, G. T. (2017), *Defining creativity*, in Plucker (ed.), "Creativity and Innovation: Theory, research and practice" Prufrock Press Inc. 2017, ISBN 978-1-61821-595-6.Kaufman, J. C. & R. J. Sternberg (ed.), (2010), *The Cambridge handbook of creativity*, Cambridge University Press, 2010

Malhotra, N. K., D. F. Birks & P. Wills (2012), *Marketing Research. An applied approach*, Pearson Education Limited, Essex, 2012

OECD (2015), *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris, https://doi.org/10.1787/9789264239012-en.

OECD/Eurostat (2019), *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition*, The Measurement of Scientific, Technological and Innovation Activities, OECD Publishing, Paris/Eurostat, Luxembourg. Online: https://doi.org/10.1787/9789264304604-en.

Plucker, J.A. (2017), *Creativity. It's Not Just for Hippies Anymore*, in Plucker (ed.), "Creativity and Innovation: Theory, research and practice" Prufrock Press Inc. 2017, ISBN 978-1-61821-595-6.

Tidd, J., J. Bessant & K. Pavitt (2005), "*Managing Innovation Integrating Technological Market and Organizational Change*", 3rd edition, J. Wiley and Sons, ISBN 0470093269

Hvem er uddannelseshjælpsmodtagerne?

Lisbeth Palmhøj Nielsen, Chris Cornelia Friis Christiansen og Anna Hansen
Center for Viden og Analyse
Styrelsen for Arbejdsmarked og Rekruttering

Sammenfatning

Formålet med analysen er at belyse, hvad der karakteriserer forskellige typer af uddannelseshjælpsmodtagere. En større viden om uddannelseshjælpsmodtagerne kan understøtte en målrettet indsats, så de unge hurtigst muligt hjælpes i uddannelse eller job. Mens nogle har brug for et venligt skub ud af kontanthjælpssystemet, hvorefter de aldrig vender tilbage, har andre brug for en mere håndholdt overgang for sikre fastholdelse i uddannelse og job.

Fra et statistisk perspektiv tegner der sig et billede af fire forskellige typer af uddannelseshjælpsmodtagere i analysen. De fire typer findes inden for alle tre visitationskategorier, dog med gradvist flere og mere komplekse barrierer for uddannelse og job fra de åbenlyst uddannelsesparate over de uddannelsesparate og til de aktivitetsparate.

Typerne af uddannelseshjælpsmodtagere i analysen dækker over:

- *Ingen eller få åbenlyse barrierer for uddannelse og job*
  Store grupper blandt de åbenlyst uddannelsesparate (omkring halvdelen) og uddannelsesparate (knap 40 pct.) har ikke åbenlyse barrierer for at komme i uddannelse og job. De ligger generelt under gennemsnittet for kriminalitet og brug af sundhedsydelser, og deres forældre er hovedsageligt beskæftigede. Mens en stor andel har uddannelsesfrafald bag sig, er der også flere med en gymnasial uddannelse. For gruppen med gymnasial uddannelse gælder det på tværs af visitationskategorier, at de har en relativt høj afgang fra kontanthjælpssystemet og også i høj grad finder varigt fodfæste uden for systemet.
- *Mødre*
  Kvinder med børn skiller sig ud ved at have en betydeligt længere anciennitet i kontanthjælpssystemet (gennemsnitligt mellem 1-4 år) end andre i samme visitationskategori. Omkring halvdelen af klyngen bestående af mødre er enlige forsørgere. Inden for alle visitationskategorier observeres det samme mønster; at mødrene har lav afgang fra kontanthjælpssystemet, men for dem som afgår, fastholdes de i uddannelse og job på niveau med deres visitationskategori.
- *Mænd med forældre langt fra arbejdsmarkedet*
  En stor gruppe mænd har forældre med meget lave beskæftigelsesgrader. En overvægt af gruppen har anden etnisk baggrund end dansk, og en stor andel er tidligere anbragte. Særligt blandt uddannelsesparate og aktivitetsparate forekommer også en overhyppighed af kriminalitet, hvor omkring halvdelen af gruppen har været anholdt og hver fjerde har siddet i fængsel. Gruppen afgår fra kontanthjælpssystemet på samme niveau som inden for visitationskategorierne, men blandt de som afgår, er der en stor andel som *ikke* finder varigt fodfæste i uddannelse og job

- *Udfordringer med sygdom*
  Inden for alle tre visitationskategorier optræder en mindre gruppe, som har udfordringer med fysisk eller psykisk sygdom. For de åbenlyst uddannelsesparate er der en gruppe af tidligere sygedagpengemodtagere, som ellers ikke har andre sygdomstegn. Blandt de uddannelsesparate er der en meget lille klynge (1 pct.), som har et stort forbrug af sundhedsydelser, herunder hospitalsindlæggelser, lægebesøg, psykolog- og psykiaterbesøg. På tværs af de uddannelsesparate og aktivitetsparate er der en gruppe personer med psykiske udfordringer. En stor andel af denne gruppe har en gymnasial uddannelse og forældre med høj beskæftigelse og høje uddannelser. På trods af psykiske udfordringer har personer med gymnasial uddannelse en høj afgang fra kontanthjælpssystemet og senere fastholdelse i uddannelse og job. Modsat ses det for gruppen med sygdomsudfordringer, som *ikke* har en uddannelse, at de har den laveste afgang fra kontanthjælpssystemet overhovedet og i tillæg hertil også lav efterfølgende fastholdelse i uddannelse og job.

1. Indledning

Unge under 30 år uden en erhvervskompetencegivende uddannelse, der i en periode ikke kan forsørge sig selv, kan anmode kommunen om uddannelseshjælp. I januar 2019 var der ca. 36.100 personer der modtog uddannelseshjælp, hvilket svarer til 3,9 pct. af befolkningen mellem 18-29 år[1].

Formålet med analysen er at få en mere helhedsorienteret forståelse af, hvem personer som modtager uddannelseshjælp er, for at kunne understøtte deres vej i uddannelse og job på bedst mulig vis.

På baggrund af en række karakteristika (demografi, højest fuldførte uddannelse, sociale karakteristika, kontakt til sundhedsvæsen og uddannelses- og beskæftigelseshistorik) inddeles uddannelseshjælpsmodtagere i mindre grupper, kaldet klynger, alt efter hvor ensartede personernes karakteristika er. Det er disse klynger, som beskrives nærmere i analysen. I analysen inddrages også data om afgangsmønstre.

[1] Jobindsats.dk, Danmarks Statistik (FOLK1A) og egne beregninger.

Metode

Analysen foretages ved hjælp af en såkaldt klyngeanalyse. En klyngeanalyse er en statistisk metode, der anvendes til at inddele karakteristika i klynger baseret på interne ligheder. I dette tilfælde er der således sket en mekanisk inddeling af uddannelseshjælpsmodtagerne på baggrund af hvor ensartede deres baggrundskarakteristika er. På den måde får man en helhedsorienteret indsigt i typer af personer i uddannelseshjælp, og kommer på den måde et spadestik dybere end gennemsnitsbetragtninger.

Klyngeanalysen er foretaget separat for de tre visitationskategorier, der overordnet arbejdes med i forbindelse med uddannelseshjælpsmodtagere: åbenlyst uddannelsesparat, uddannelsesparat og aktivitetsparat. Analysens population består af alle personer der modtog uddannelseshjælp i løbet af 2016 (jf. bilag 1).

Til klyngedannelsen er der anvendt mange forskellige karakteristika inden for demografi, højest fulførte uddannelse, sociale, kontakt til sundhedsvæsen, samt forudgående uddannelses- og beskæftigelseshistorik.

Det skal bemærkes at metoden er følsom over for hvilke variable, der anvendes. De identificerede klynger er derfor et produkt af det tilgængelige datamateriale. Med et andet datagrundlag ville der ikke nødvendigvis fremkomme samme inddeling. Metode og fremgangsmåde er nærmere beskrevet i bilag 1 og 2.

2. Størstedelen af de åbenlyst uddannelsesparate har få eller ingen åbenlyse barrierer for uddannelse og job

Uddannelseshjælpsmodtagere visiteres åbenlyst uddannelsesparate, hvis de vurderes ikke at have barrierer for at kunne starte på en uddannelse og gennemføre den på ordinære vilkår. Ud af de tre visitationskategorier vurderes de åbenlyst uddannelsesparate at have færrest barrierer for at komme i uddannelse og job. I 2016 var der knap 9.200 personer, som var visiteret åbenlyst uddannelsesparate. Det svarer til 14 pct. af den samlede gruppe uddannelseshjælpsmodtagere.

I det følgende forsøges at belyse, hvad der karakteriserer de åbenlyst uddannelsesparate, og hvilke typer der findes blandt uddannelseshjælpsmodtagere.

De åbenlyst uddannelsesparate inddeles mekanisk i fem klynger, hvor de fem klynger kan samles i tre hovedgrupper ud fra deres åbenbare barrierer for at gå i uddannelse eller job.

De baggrundskarakteristika som ligger til grund for beskrivelserne af hver klynge inden for gruppen af de åbenlyst uddannelsesparate er opgjort i bilag 3.

De to første klynger udgør tilsammen 54 pct. af de åbenlyst uddannelsesparate og skiller sig ud ved *ikke* at have nogen åbenlyse barrierer for uddannelse og job. Begge klynger lader til at kunne gå i uddannelse og job ved at få et lille skub.

1. Unge med dansk herkomst uden uddannelse (32 pct.)

Den første klynge består af yngre personer (gennemsnitligt 21 år) med dansk herkomst. Ingen i klyngen har fuldført højere uddannelse end folkeskolen. Personerne er herudover kendetegnet ved ikke at have nogen børn, og sammenlignet med andre uddannelseshjælpsmodtagere har de kun i lav grad kontakt til sundhedsvæsnet, retsvæsnet eller været anbragte som børn. Herudover har personer i gruppen typisk forældre med mellemlange uddannelser og en høj beskæftigelsesgrad. Omkring 90 pct. har modtaget SU inden for de seneste tre år forud for uddannelseshjælpsforløbet, hvilket kan tyde på, at de fleste i gruppen kommer ind i kontanthjælpssystemet med uddannelsesfrafald i baggagen. Det er dog også kendetegnene for åbenlyst uddannelsesparate generelt.

2. Unge med gymnasial uddannelse (22 pct.)

Den anden klynge repræsenterer unge, som har en gymnasial uddannelse (97 pct.). Næsten ingen i klyngen har børn og kun få har været i kontakt med rets- og sundhedsvæsenet. Ligeledes er kun få tidligere anbragte. Deres forældre har overvejende lange uddannelser og høj beskæftigelsesgrader. Som det er gældende for mange uddannelseshjælpsmodtagere har flere i denne gruppe haft kontakt til psykolog inden for de seneste fem år (10 pct.).

De næste tre klynger har karakteristika, som erfaringsmæssigt kan virke som barrierer i forhold til at påbegynde og gennemføre en uddannelse. Tilsammen udgør de tre klynger 46 pct. af de åbenlyst uddannelsesparate.

3. *Mænd med forældre med lav beskæftigelse (30 pct.)*

I denne klynge udgør mænd to ud af tre. Sammenlignet med andre åbenlyst uddannelsesparate har gruppens forældre den laveste uddannelsesgrad og de har en lav beskæftigelsesgrad. Hver fjerde har anden etnisk herkomst end dansk (gennemsnittet for åbenlyst uddannelsesparate med anden etnisk baggrund end dansk er 13 pct.). De fleste er enlige, de har ingen børn., og der er flere tidligere anbragte (16 pct.) og en større andel har været anholdt (24 pct.) og siddet i fængsel (8 pct.) i denne klynge end i de andre klynger for åbenlyst uddannelsesparate. Alle i klyngen har folkeskolen som højest gennemførte uddannelse.

4. Enlige mødre (10 pct.)

I denne klynge er alle forældre (93 pct. er mødre), hvoraf knap halvdelen er enlige forsørgere. De er i gennemsnit 25 år, hvilket er højt sammenlignet med den samlede gruppe uddannelsesmodtagere. De har tidligere været i uddannelse, og der er også en del, som har gennemført en gymnasial uddannelse (13 pct.). De har en høj grad af indlæggelser og lægebesøg, hvilket kan være som følge af graviditet og fødsel, mens de ikke har hyppigere kontakt til psykiater og psykolog end den gennemsnitlige uddannelseshjælpsmodtager. Sammenlignet med andre der er visiteret åbenlyst uddannelsesparate, har denne gruppe en betydelig længere anciennitet i kontanthjælpssystemet (1 år).

5. *Tidligere sygedagpengemodtagere (6 pct.)*

Denne klynge består hovedsageligt af tidligere sygedagpengemodtagere (84 pct.). De er ældre end gennemsnittet (24 år gamle), flere har børn (30 pct.), flere har været anbragte som børn (15 pct.), og flere har været anholdt (21 pct.) og siddet i fængsel (7 pct.). På trods af at de har en historik, hvor mange har modtaget sygedagpenge 3 år forud for uddannelseshjælp (84 pct.), har de *ikke* flere kontakter til sundhedsvæsnet end gennemsnittet.

2.1 Åbenlyst uddannelsesparate med få eller ingen åbenlyse barrierer finder oftere fodfæste

I det følgende sammenlignes afgangsmønstrene for de fem klynger. Det sker for at få et klarere billede af, om nogle klynger har større udfordringer end andre med at finde fodfæste i uddannelse og job.

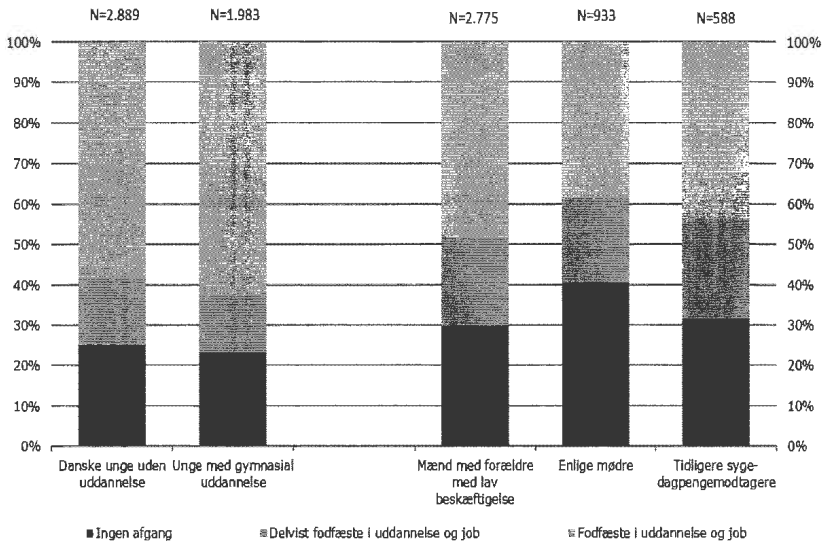
I figur 1 vises graden af hver enkelt klynges efterfølgende fodfæste i uddannelse og job. Grå afspejler, om de får fodfæste i uddannelse eller job i et år efter de afgår fra uddannelseshjælp. Rosa illustrerer unge, som afgår fra uddannelseshjælp og som enten vender tilbage til kontanthjælpssystemet inden for et år eller er i selvforsørgelse et helt år. Blå belyser de unge, der ikke afgik i hele 2016.

Den første gruppering, der består af to klynger, som lader til at have få eller ingen åbenbare barrierer for at få uddannelse og job, har som forventet større fodfæste end den anden gruppering med lidt flere barrierer. Særligt klyngen med gymnasiale uddannelser klarer sig godt. Det bemærkes, at selvom de to klynger ikke har åbenbare barrierer for uddannelse og job, er de i gennemsnit henholdsvis 4 og 6 måneder i kontanthjælpssystemet, jf. bilag 3.

Grupperingen med lidt flere barrierer har sværere ved at finde fodfæste, både helt og delvist. Særligt klyngen bestående af mødre afgår i mindre grad end de resterende klynger, men det lader dog til, at når de afgår, finder de i høj grad varigt fodfæste. Klyngerne bestående af mænd med forældre med lav beskæftigelse og tidligere sygedagpengemodtagere derimod afgår oftere, men en del vender også tilbage til kontanthjælpssystemet igen inden for et år.

Figur 1.

Afgangsmønstre – åbenlyst uddannelsesparate



Anm.: Definition af ingen afgang, delvist fodfæste i uddannelse og job, samt fodfæste i uddannelse og job er givet i bilag 2.

Kilde: DREAM

3. Uddannelsesparate har oftere komplekse barrierer for uddannelse og job

Uddannelseshjælpsmodtagere visiteres uddannelsesparate, hvis det vurderes, at de med den rette hjælp og støtte kan påbegynde en uddannelse inden for et år. I 2016 var der knap 32.600 uddannelsesparate uddannelseshjælpsmodtagere, hvilket svarer til 49 pct. af den samlede gruppe uddannelseshjælpsmodtagere. Det er kendetegnende for gruppen, at størstedelen har folkeskolen som højest gennemførte uddannelse (89 pct.), ligesom størstedelen har perioder med SU forud for ydelsesforløbets start (72 pct.). Det kan vidne om, at mange i denne visitationsgruppe har afbrudte uddannelsesforløb bag sig. Sammenlignet med gruppen af åbenlyst uddannelsesparate, har de uddannelsesparate flere karakteristika, der kan virke som barrierer for at komme i uddannelse og job, eksempelvis har flere været anbragt som børn (18 pct.), og en større andel har været anholdt inden for de seneste 10 år (22 pct.).

Når de uddannelsesparate mekanisk inddeles i klynger, fremkommer igen fem klynger, som kan samles i tre hovedgrupper baseret på graden af barrierer for at komme i uddannelse og job. Baggrundskarakteristika for hver klynge er opgjort i bilag 3.

1. *Unge med dansk herkomst uden uddannelse (39 pct.)*

Den første gruppe består af en enkelt klynge. Ligesom den største klynge blandt åbenlyst uddannelsesparate, adskiller gruppen sig ved *ikke* at have åbenlyse hindringer for at gå i uddannelse eller job. Klyngen består af unge personer (gennemsnitligt 21 år) med dansk herkomst, hvis forældre er uddannede og har høje beskæftigelsesgrader. Selv har alle de unge i klyngen folkeskolen som højest gennemførte uddannelse. Sammenlignet med den restende gruppe uddannelsesparate har de en kort anciennitet i kontanthjælpssystemet (0,9 år).

Den næste gruppe består af to klynger, som lader til at have brug for lidt mere støtte til at komme i uddannelse eller beskæftigelse. Gruppen udgør halvdelen af alle uddannelsesparate.

2. *Tidligere kriminelle og anbragte mænd med forældre med lav beskæftigelse (38 pct.)*

Den første klynge består hovedsageligt af mænd (73 pct.), hvor hver fjerde har anden etnisk herkomst end dansk. Personerne er yderligere kendetegnet ved, at forældre hovedsageligt er ufaglærte med lave beskæftigelsesgrader (21-23 pct.). Hver tredje er tidligere anbragte. Knap halvdelen af personerne i klyngen har været anholdt, hver femte har siddet i fængsel.

3. *Enlige mødre med forældre med lav beskæftigelse (12 pct.)*

Den anden klynge i gruppen, som ser ud til at have brug for lidt støtte for at gå i uddannelse og job, er alle forældre og de flest er kvinder (91 pct.). Halvdelen er også enlige forsørgere. De er i gennemsnit er 25 år, hvilket er højt ift. den resterende gruppe uddannelseshjælpsmodtagere. Omkring halvdelen af deres forældre har ingen uddannelse og lave beskæftigelsesgrader (30-34 pct.). De bruger i høj grad sundhedsvæsenet, både i form af lægekontakter, indlæggelser og ambulante behandlinger. Den høje kontakt til sundhedsvæsenet kan potentielt relateres til graviditet og fødsel. Personerne i denne klynge har en relativ lang anciennitet i kontanthjælpssystemet (1,7 år).

De sidste to klynger, som udgør 12 pct. af alle uddannelsesparate, lader til at være udfordret af sygdom.

4. *Personer med gymnasial uddannelse og psykiske udfordringer (11 pct.)*

Den første af de to klynger består af personer, der i gennemsnit er 24 år, og som har gennemført en gymnasial uddannelse (84 pct.) En stor andel har kontakt til psykolog (27 pct.). Sammenlignet med andre uddannelsesparate har de relativ kort anciennitet i kontanthjælpssystemet (1 år), og flere har tidligere modtaget sygedagpenge (18 pct.). Blandt deres forældre er der en overrepræsentation med videregående uddannelser (39 pct.) og høje beskæftigelsesgrader (55-57 pct.).

5. *Kvinder med både fysiske og psykiske udfordringer (1 pct.)*

Den anden klynge, som kun svarer til 1 pct. af alle uddannelsesparate, består af kvinder (86 pct.) der har betydeligt større kontakt til sundhedsvæsenet end andre uddannelseshjælpsmodtagere, både i form af kontakt til læge, hospitalsindlæggelser, ambulante behandlinger, og kontakt til psykolog og psykiater. Hver fjerde i denne lille klynge er tidligere anbragte, og er deslige kendetegnet ved at have en lang anciennitet i kontanthjælpssystemet (2,1 år).

3.1 Personer med gymnasial uddannelse finder fodfæste i uddannelse og job

Afgangen til uddannelse og job for de uddannelsesparates klynger er mere differentierede end for de åbenlyst uddannelsesparate. For de uddannelsesparate er klyngerne mere komplekse, og hver klynge har mere forskelligartede barrierer, og det afspejles i afgangsmønstrene illustreret i figur 3.

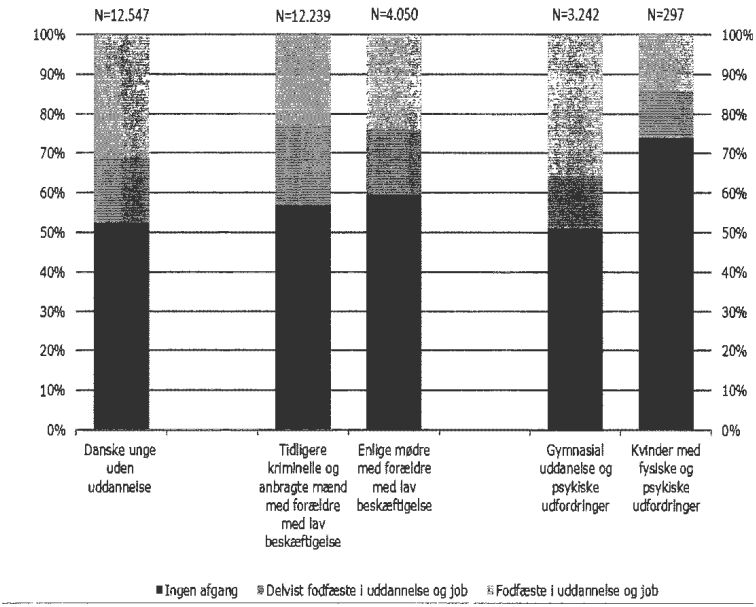
Ses der bort fra den lille klynge på 1 pct., ser man at selvom klyngerne har næsten lige stor afgang fra kontanthjælpssystemet (mellem 41 og 50 pct.), er der stor spredning på graden af efterfølgende fodfæste i uddannelse og job. Som man så det for de åbenlyst uddannelsesparate har personer med gymnasial uddannelse både størst afgang fra kontanthjælpssystemet, og også efterfølgende bedst fodfæste i uddannelse og job. På trods af psykiske udfordringer opnår de også mere fodfæste end den klynge, der vurderes at have få eller ingen åbenlyse barrierer for uddannelse og job.

Ses der igen bort fra klyngen som udgør 1 pct., har klyngen af uddannelsesparate mødre ligesom klyngen af mødre blandt de åbenlyst uddannelsesparate, den laveste afgang fra systemet, men dem som afgår, finder i høj grad varigt fodfæste. Modsat ser det ud for klyngen bestående af tidligere anbragte mænd med forældre med lave beskæftigelsesgrader. Selvom de afgår fra kontanthjælpssystemet ligesom gennemsnittet af uddannelsesparate, har de sværest ved at forblive i uddannelse og job, og kun halvdelen, af de som afgår, får varigt fodfæste inden for det fulgte år.

Af figur 3 fremgår det deslige, at kvinder med fysiske og psykiske problemer har en lavere afgang fra kontanthjælpssystemet end de resterende klynger. Klyngen er dog lille, kun 1 pct. af alle uddannelsesparate.

Figur 3.

Afgangsmønstre – uddannelsesparate



Anm.: Definition af ingen afgang, delvist fodfæste i uddannelse og job, samt fodfæste i uddannelse og job er givet i bilag 1.
Kilde: DREAM

4. Aktivitetsparate træder tidligt ind i kontanthjælpssystemet

Uddannelseshjælpsmodtagere visiteres aktivitetsparate, hvis de vurderes at have brug for mere omfattende hjælp og støtte i længere tid end ét år, inden de kan påbegynde en uddannelse, som de kan gennemføre på almindelige vilkår. Der var knap 24.300 aktivitetsparate i 2016, hvilket svarer til 37 pct. af alle uddannelseshjælpsmodtagerne.

Gruppen af aktivitetsparate er generelt kendetegnet ved at en stor andel har haft kontakt til sundheds- og retsvæsnet. Hver fjerde er også tidligere anbragt (27 pct.). Til trods for at gennemsnitsalderen er 22 år, har de aktivitetsparate i gennemsnit en anciennitet på 2,8 år i kontanthjælpssystemet, hvilket betyder at en stor andel af de aktivitetsparate har ansøgt om uddannelseshjælp, så snart de er fyldt 18 år.

Endnu engang inddeles gruppen i visitationskategorien mekanisk i klynger, og igen dannes fem klynger, som også her kan samles i tre hovedgrupper. Bilag 5 viser baggrundkarakteristika for hver klynge blandt de aktivitetsparate.

1. Unge mænd med dansk herkomst uden uddannelse (33 pct.)
Den første klynge består af mænd (77 pct.) med dansk herkomst (97 pct.), som er forholdsvis unge (21 år). Ingen i klyngen har anden uddannelse end folkeskolen. De har relativet lang anciennitet i kontanthjælpssystemet givet deres alder (1,7 år) og kun få i klyngen har perioder med beskæftigelse eller SU forud for tilkendelsen af uddannelseshjælp. 23 pct. af gruppen har været anbragte som børn. Herudover er der ikke udslagsgivende karakteristika, der kan bidrage med en forklaring på, hvorfor de unge i denne klynge ikke er i uddannelse eller job eller hvorfor de er visiteret aktivitetsparate.

De næste to klynger tilhører en gruppering, som har en række karakteristika, som kan udgøre betydelige barrierer for uddannelse og job.

2. Tidligere anbragte mødre med forældre med lav beskæftigelse (13 pct.)
Den første klynge består hovedsageligt af forældre (84 pct.), hvor de fleste er kvinder (89 pct.) og ældre end gennemsnittet (24 år). En stor del af gruppen har været anbragt som barn (34 pct.), og har forældre med lave beskæftigelsesgrader (26-30 pct.). Hver femte har anden etnisk herkomst end dansk. De har høj anciennitet i kontanthjælpssystemet (i gennemsnit 4 år). De har mange kontakter til sundhedsvæsenet i form af indlæggelser, ambulante behandlinger og lægekontakter, som potentielt kan henføres til graviditet og fødsel, mens de sammenlignet med hele gruppen af aktivitetsparate ikke har markant større kontakt til psykiatrien.

3. Tidligere kriminelle og anbragte mænd med forældre med meget lav beskæftigelse (12 pct.)
Den anden klynge består af enlige (74 pct.) mænd (83 pct.) med en overvægt af unge med anden etnisk herkomst end dansk (61 pct.). Omkring halvdelen af klyngen har været anbragt som barn. Deres forældre har et lavt uddannelsesniveau og meget lav beskæftigelsesgrad (13-14 pct.). I klyngen har 62 pct. været anholdt, mens 44 pct. har siddet i fængsel.

Den sidste gruppe består af to klynger, som begge er karakteriseret ved sygdomstegn og repræsenterer 41 pct. af alle aktivitetsparate.

4. Tidligere anbragte enlige med psykiske udfordringer og lang anciennitet i kontanthjælpssystemet (23 pct.)
I den første klynge er en tredjedel tidligere anbragte, fire af fem er enlige og gennemsnittet for anciennitet i kontanthjælpssystemet er 5,2 år, hvilket svarer til hele deres voksne liv siden de fyldte 18 år – de er i gennemsnit 23 år gamle. Desuden har de mange kontakter til læge, psykiater og psykolog.

5. Kvinder med psykiske udfordringer, gymnasial uddannelse og højtuddannede forældre (18 pct.)
I den anden klynge er 60 pct. kvinder, de har dansk herkomst (97 pct.), og en stor andel har en gymnasial uddannelse (25 pct.). En stor andel af dem (ca. 30 pct.) har været til psykiater eller psykolog inden for det seneste år. Deres forældre er veluddannede (kun 7 pct. har folkeskolen som højest fuldførte) og har høje beskæftigelsesgrader (60-62 pct.).

4.1 Kun få aktivitetsparate får fodfæste i uddannelse og job inden for et år

Afgangsmønstrene skal ses i lyset af, at gruppen af aktivitetsparate netop er det, fordi det er vurderet, at de ikke kan starte uddannelse inden for et år.Den første klynge, som på observerbare data er kendetegnet ved få eller ingen åbenlyse barrierer for uddannelse og job, har den næstlaveste afgang fra kontanthjælpssystemet af de fem klynger. Kun 15 pct. afgik fra kontanthjælpssystemet i 2016, jf. figur 5. Blandt dem som afgår, er det kun halvdelen som finder varigt fodfæste. Blandt de som afgår er det omkring halvdelen som finder varigt fodfæste.

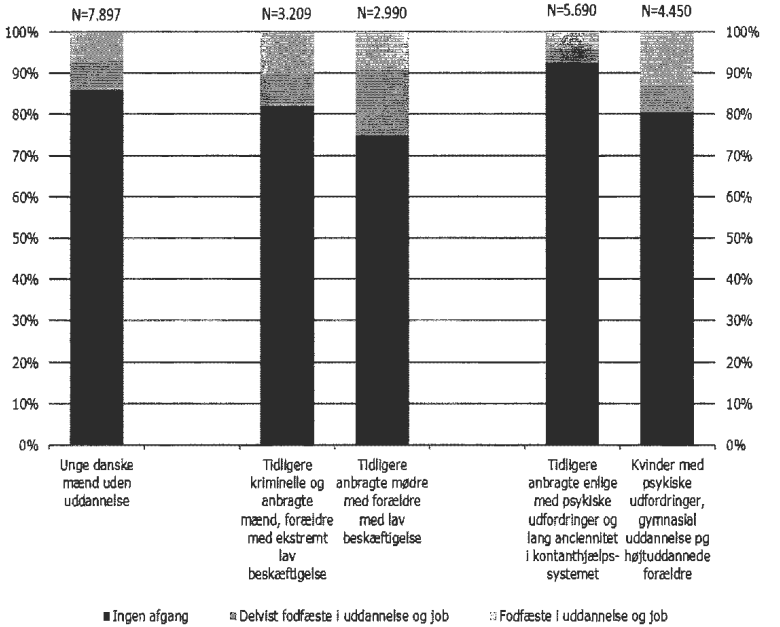
Som for de åbenlyst uddannelsesparate og uddannelsesparate har klyngen, hvor en stor del har gennemført en gymnasial uddannelse både højst afgang fra kontanthjælpssystemet og også det bedste efterfølgende fodfæste til trods for psykiske udfordringer. Niveauet for afgang er dog noget lavere sammenlignet med de andre visitationskategorier.

Som det også var tilfældet for åbenlyst uddannelsesparate og uddannelsesparate har mødre den laveste afgang fra kontanthjælpssystemet. Ligeledes har klyngen af tidligere kriminelle og anbragte mænd med forældre med ekstremt lave beskæftigelsesgrader på trods af højere afgang mindre varigt fodfæste.

Klyngen med psykiske udfordringer, uden en gymnasial uddannelse, har den laveste afgang fra kontanthjælpssystemet, hvor kun 8 pct. afgår. Og blandt de som afgår, er det kun omkring halvdelen som finder varigt fodfæste.

Figur 5.

Afgangsmønstre – aktivitetsparate



Anm.: Definition af ingen afgang, delvist fodfæste i uddannelse og job, samt fodfæste i uddannelse og job er givet i bilag 3.

Kilde: DREAM

Bilag 1: Data og definitioner
I det følgende fremgår en kort beskrivelse af data og centrale definitioner.

Data og population
I analysen er anvendt samme population som i STARs tidligere analyse af uddannelseshjælpsmodtagere: Unges vej fra uddannelseshjælp til uddannelse og beskæftigelse. Her benyttes beskæftigelsesministeriets forløbsdatabase, DREAM, til at udtrække populationen.

Analysens population består af alle personer, der er tilgået et uddannelseshjælpsforløb siden januar 2014, hvor uddannelseshjælp blev indført, og som har modtaget uddannelseshjælp i løbet af 2016. Gruppen, der indgår i analysen, består således både af unge, der har været på uddannelseshjælp i længere tid, samt unge, der tilgår uddannelseshjælp i løbet af 2016.

DREAM er anvendt til at beskrive afgangsmønstre og i kombination med mentordata også indsatsmønstre. Der er anvendt Danmarks Statistiks befolkningsdata, familiedata, sundhedsdata, kriminalitetsstatistikker, samt opgørelser over anbringelser og forebyggende statistikker.

Definition af afgangsmønster
Et forløb er i analysen afsluttet, når personen i fire sammenhængende uger ikke har modtaget uddannelseshjælp. Den nye status personen afgår til, er givet ved et nedslag i den femte uge efter afgangen fra uddannelseshjælp.

Populationen inddeles i tre grupper, alt efter hvor tæt de er på en uddannelse eller arbejdsmarkedet.

Fodfæste i uddannelse eller job
- Gruppen defineres ved personer, der er afgået fra uddannelseshjælp i 2016 og har haft mindst én uges uddannelse eller beskæftigelse i det efterfølgende år, og samtidig ikke på noget tidspunkt vender tilbage til ydelsessystemet i samme tidsperiode.

Delvist fodfæste uden for ydelsessystemet
- Personer der er afgået til selvforsørgelse i 2016 og forbliver selvforsørgende i det efterfølgende år[2].
- Personer der i 2016 afgår til enten job, uddannelse eller selvforsørgelse, men som inden for et år vender tilbage til ydelsessystemet.

Fortsat i ydelsessystemet
- Personer der *ikke* afgår fra uddannelseshjælp i 2016.
- Personer der afgår direkte til kontanthjælp i 2016.
- Personer der afgår direkte til anden ydelse i 2016.

[2] Personer i fængsel får ingen ydelse, hvorfor afgang til selvforsørgelse kan skyldes fængsling.

Bilag 2: Metode og fremgangsmåde

Klyngeanalyse (eller cluster analysis på engelsk) er en fællesbetegnelse for en række statistiske metoder eller teknikker til gruppering af observationer, i dette tilfælde gruppering af personer.

I denne analyse er der indledningsvist anvendt en hierarkisk clustering metode til at bestemme antallet af klynger. Herefter er der foretaget en ikke-hierarkisk clustering metode til mere korrekt at inddele observationerne i klynger og tildele personer til en bestemt klynge. Proceduren er udført separat for åbenlyst uddannelsesparate, uddannelsesparate og aktivitetsparate. Populationen for klyngeanalysen er hhv. 9.200 åbenlyst uddannelsesparate, 32.600 uddannelsesparate og 24.200 aktivitetsparate uddannelseshjælpsmodtagere.

Der er anvendt følgende baggrundskarakteristika til gruppering af personerne:

Demografiske karakteristika:

- Alder
- Køn
- Herkomst
- Hvorvidt man er enlig
- Hvorvidt man har børn
- Hvorvidt man er forsørger
- Hvorvidt man er enlig forsørger

Højest fuldført uddannelse:

- Folkeskolen, gymnasial uddannelse, erhvervsuddannelse eller videregående uddannelse

Sociale karakteristika:

- Forældrenes højest fuldførte uddannelse: Folkeskole, gymnasial uddannelse, erhvervsuddannelse eller videregående uddannelse
- Moderens beskæftigelsesgrad siden 2008
- Faderens beskæftigelsesgrad siden 2008
- Hvorvidt man har været anbragt som barn
- Hvorvidt man har været fængslet inden for de seneste 10 år
- Hvorvidt man har været anholdt inden for de seneste 10 år

Kontakt til sundhedsvæsen:

- Hvorvidt man har været indlagt på hospitalet inden for de seneste 5 år
- Antal lægebesøg inden for det seneste år
- Hvorvidt man har været til privat psykiater inden for de seneste 5 år
- Hvorvidt man har været til privat psykolog inden for de seneste 5 år
- Hvorvidt man har modtaget ambulante behandlinger inden for de seneste 5 år

Forudgående uddannelses- og beskæftigelseshistorik

- Hvorvidt man har været beskæftiget 3 år forud for ydelsesforløbets start
- Hvorvidt man har modtaget sygedagpenge 3 år forud for ydelsesforløbets start
- Hvorvidt man har modtaget SU 3 år forud for ydelsesforløbets start
- Anciennitet i kontanthjælpssystemet

Den hierarkiske clustering metode fungerer ved, at hver person i udgangspunktet indgår i sin egen klynge, og processen er sådan, at de to tætteste klynger (dem der ligner personen mest) samles til en ny klynge, som erstatter de to forrige klynger. Iterationsprocessen fortsætter til der kun er én klynge.

Man kan bruge forskellige metoder til at finde ud af, hvor ens to personer er (eller hvor stor afstanden er mellem observationer), og i dette tilfælde er der anvendt Wards Minimum. Ved denne metode samles de to klynger, hvor clustering metoden skaber den mindste stigning i sum of the squared i klyngen.

De seneste 15 eller 20 trin i clustering processen bestemmer, hvor mange klynger der meningsfyldt skal være. I dette tilfælde viste den hierarkiske clustering metode en løsning med hhv. fem klynger for hver visitationskategori. Det er tilfældigt at antallet er fem for alle tre visitationskategorier.

Efter at have fået overblik over outliers og antallet af klynger, blev den ikke-hierarkiske clusteranalyse gennemført. Denne metode er fordelagtig til den endelige inddeling af observationer og dannelse af klyngerne, da den tillader at observationerne kan hoppe mellem klyngerne, mens de dannes. Her anvendes de første 10 observationer som kerner i hver af deres egne fem klynger. De efterfølgende observationer tildeles så de fem klynger på baggrund af afstandene til klyngernes kerner. De oprindelige ti kerner er ikke faste, men kan som alle observationer flytte klynge alt efter, hvem de ligner mest. Processen fortsætter indtil hver klynge har en kerne, der er lig gennemsnittet for klyngen på de anvendte baggrundskarakteristika.

Herudover skal det bemærkes at klyngeanalyse er meget følsom over for kombinationen af de baggrundskarakteristika, som anvendes i grupperingen af uddannelseshjælpsmodtagere, og den måde man opgør variablene på. Klyngerne er derfor et resultat af det tilgængelige data, de udvalgte baggrundskarakteristika og måden de er bearbejdet og sorteret.

Bilag 3

Karakteristika for åbenlyst uddannelsesparate, fordelt på klynger

	Folke-skole	Gym.	Forælddre lav besk.	Mødre	Tidl. syge-dp.	ÅUP	Alle uddhj.
Demografiske karakteristika							
Alder (gns.)	21	23	22	25	24	22	22
Andel mænd	51	59	66	7	58	53	54
Andel med anden etnisk baggrund	3	14	25	14	13	13	14
Andel enlige	43	63	73	46	51	57	63
Andel med børn	0	3	2	100	30	14	15
Andel forsørgere*	0	3	0	99	18	12	12
Andel enlige forsørgere*	0	2	0	45	0	5	5
Højst fuldførte uddannelse (andel)							
Folkeskolen	100	2	100	86	93	77	89
Gymnasiale uddannelser	0	97	0	13	4	23	10
Erhvervsuddannelser	0	1	0	1	3	2	1
Videregående uddannelser	0	0	0	0	0	0	0
Sociale karakteristika							
Forældrenes højst fuldførte uddannelse (andel)							
- Folkeskolen	7	12	43	34	37	24	31
- Gymnasiale uddannelser	5	7	2	0	1	4	3
- Erhvervsuddannelser	57	39	38	50	45	46	44
- Videregående uddannelser	30	41	17	16	17	26	22
Moderens beskæftigelsesgrad siden 2008	71	61	24	40	39	49	42
Faderens beskæftigelsesgrad siden 2008	73	59	25	41	42	50	44
Andel anbragt som barn	2	3	16	14	15	9	20
Andel som har siddet i fængsel de seneste 10 år	0	1	8	2	7	3	9
Andel som har været anholdt de seneste 10 år	4	7	24	9	21	12	21
Kontakt til sundhedsvæsenet							
Andel med hospitalsindlæggelse seneste 5 år**	8	11	14	69	19	17	20
Antal lægebesøg seneste år (gns.)	5,4	5,3	5,2	10,4	7,1	5,9	7,6
Andel med kontakt til privat psykiater seneste 5 år	3	3	3	5	6	3	10
Andel med kontakt til psykolog seneste 5 år	11	10	3	10	7	8	11
Andel med ambulante behandlinger seneste 5 år	45	46	52	95	61	53	59
Uddannelses- og beskæftigelseshistorik							
Andel med beskæftigelse 3 år forud for ydelsesforløbets start	77	71	57	46	92	67	51
Andel som har modtaget sygedagpenge 3 år forud for ydelsesforløbets start	1	8	0	9	84	8	7
Andel som har modtaget SU 3 år forud for ydelsesforløbets start	90	95	84	89	72	88	62
Anciennitet i kontanthjælpssystemet, år (gns.)	0,4	0,5	0,6	1	0,5	0,5	1,7
Personer i alt	2.889	1.983	2.775	933	588	9.168	65.961

Anm.: *Forsørgere dækker over personer med hjemmeboende ugifte børn under 25 år. ** På somatiske afdelinger.
Kilde: DREAM samt Danmarks Statistiks befolkningsdata, familiedata, sundhedsdata, kriminalitetsstatistik samt data vedr. unges anbringelser og egne beregninger.

Bilag 4

Karakteristika for uddannelsesparate, fordelt på klynger

	Folke-skole	Forældre lav besk.	Mødre	Gym. psyk.	psyk. fys.	UP	Alle uudbj.
Demografiske karakteristika							
Alder (gns.)	21	23	25	24	22	22	22
Andel mænd	49	73	9	51	14	53	54
Andel med anden etnisk baggrund	4	26	19	11	9	15	14
Andel enlige	52	74	53	66	62	62	63
Andel med børn	1	8	100	3	8	16	15
Andel forsørgere*	0	3	93	3	3	13	12
Andel enlige forsørgere*	0	0	48	0	0	6	5
Højst fuldførte uddannelse (andel)							
- *Folkeskolen*	100	99	95	9	94	89	89
- *Gymnasiale uddannelser*	0	0	4	88	5	10	10
- *Erhvervsuddannelser*	0	0	1	3	1	1	1
- *Videregående uddannelser*	0	0	0	0	0	0	0
Sociale karakteristika							
Forældrenes højst fuldførte uddannelse (andel)							
- *Folkeskolen*	6	57	45	15	29	31	31
- *Gymnasiale uddannelser*	4	2	2	4	3	3	3
- *Erhvervsuddannelser*	62	28	40	40	46	44	44
- *Videregående uddannelser*	26	12	12	39	21	21	22
Moderens beskæftigelsesgrad siden 2008	60	21	30	57	34	41	42
Faderens beskæftigelsesgrad siden 2008	64	23	34	55	50	44	44
Andel anbragt som barn	9	31	21	5	26	18	20
Andel som har siddet i fængsel de seneste 10 år	0	22	2	2	4	9	9
Andel som har været anholdt de seneste 10 år	6	46	10	9	14	22	21
Kontakt til sundhedsvæsenet							
Andel med hospitalsindlæggelse seneste 5 år**	15	11	64	16	57	20	20
Antal lægebesøg seneste år (gns.)	7,3	5,4	11,7	8,3	47,1	7,6	7,6
Andel med kontakt til privat psykiater seneste 5 år	11	6	8	11	18	9	10
Andel med kontakt til psykolog seneste 5 år	13	4	11	27	28	11	12
Andel med ambulante behandlinger seneste 5 år	57	49	93	56	95	59	59
Uddannelses- og beskæftigelseshistorik							
Andel med beskæftigelse 3 år forud for ydelsesforløbets start	59	53	36	65	60	54	51
Andel som har modtaget sygedagpenge 3 år forud for ydelsesforløbets start	5	7	9	18	12	8	7
Andel som har modtaget SU 3 år forud for ydelsesforløbets start	78	60	76	92	71	72	62
Anciennitet i kontanthjælpssystemet, år (gns.)	0,9	1,2	1,7	1,0	2,1	1,1	1,7
Personer i alt	12.547	12.239	4.050	3.424	297	32.557	65.961

Anm.: *Forsørgere dækker over personer med hjemmeboende ugifte børn under 25 år. ** På somatiske afdelinger.

Kilde: DREAM samt Danmarks Statistiks befolkningsdata, familiedata, sundhedsdata, kriminalitetsstatistik samt dat vedr. unges anbringelser og egne beregninger.

Bilag 5

Karakteristika for aktivitetsparate, fordelt på klynger

	Folke-skole	Mød-re	For-ældre lav besk.	Tidl. An-bragte psyk.	Gym. psyk.	AP	Alle ud-dhj.
Demografiske karakteristika							
Alder (gns.)	21	24	22	23	22	22	22
Andel mænd	77	11	83	52	39	56	54
Andel med anden etnisk baggrund	3	22	61	8	3	14	14
Andel enlige	63	53	75	79	66	67	63
Andel med børn	5	84	4	5	1	14	15
Andel forsørgere*	1	74	1	1	0	10	12
Andel enlige forsørgere*	0	34	0	0	0	4	5
Højst fuldførte uddannelse (andel)							
- Folkeskolen	100	96	99	97	72	94	89
- Gymnasiale uddannelser	0	2	1	2	25	6	10
- Erhvervsuddannelser	0	1	0	1	2	1	1
- Videregående uddannelser	0	0	0	0	0	0	0
Sociale karakteristika							
Forældrenes højst fuldførte uddannelse (andel)							
- *Folkeskolen*	35	50	45	31	7	32	31
- *Gymnasiale uddannelser*	0	2	16	4	1	3	3
- *Erhvervsuddannelser*	39	37	26	45	59	42	44
- *Videregående uddannelser*	24	11	12	21	32	22	22
Moderens beskæftigelsesgrad siden 2008	44	26	14	38	62	40	42
Faderens beskæftigelsesgrad siden 2008	50	30	13	42	60	43	44
Andel anbragt som barn	23	34	48	31	10	27	20
Andel som har siddet i fængsel de seneste 10 år	9	4	44	8	1	11	9
Andel som har været anholdt de seneste 10 år	20	13	62	23	6	22	21
Kontakt til sundhedsvæsenet							
Andel med hospitalsindlæggelse seneste 5 år**	8	66	11	19	19	21	20
Antal lægebesøg seneste år (gns.)	4,5	13,7	4,5	11,6	8,5	8,1	7,6
Andel med kontakt til privat psykiater seneste 5 år	3	15	10	23	29	14	10
Andel med kontakt til psykolog seneste 5 år	1	10	2	16	30	11	12
Andel med ambulante behandlinger seneste 5 år	49	94	50	66	63	62	59
Uddannelses- og beskæftigelseshistorik							
Andel med beskæftigelse 3 år forud for ydelsesforløbets start	27	36	39	45	56	39	51
Andel som har modtaget sygedagpenge 3 år forud for ydelsesforløbets start	1	12	2	11	8	6	7
Andel som har modtaget SU 3 år forud for ydelsesforløbets start	21	42	32	38	72	39	62
Anciennitet i kontanthjælpssystemet, år (gns.)	1,7	4,0	1,4	5,2	1,8	2,8	1,7
Personer i alt	7.897	3.209	2.990	5.690	4.450	24.236	65.961

Anm.: *Forsørgere dækker over personer med hjemmeboende ugifte børn under 25 år. ** På somatiske afdelinger.

Kilde: DREAM samt Danmarks Statistiks befolkningsdata, familiedata, sundhedsdata, kriminalitetsstatistik samt data vedr. unges anbringelser og egne beregninger.

Application of a spatial Difference-in-Difference approach on a Danish tax exemption reform.

Jørgen T. Lauridsen and Morten Skak

Department of Business and Economics, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark, jtl@sam.sdu.dk

Key Words: Difference-in-differences; Program evaluation; Regional development; Spatial autocorrelation; Spatial interaction

JEL Classifications: C21, J23, R58

Abstract

By applying a recently suggested spatial Difference-in-Difference approach, the present study shows that a traditional non-spatial approach seriously misestimates effects of an economic intervention in a case with spatial interdependence between treated regions. Like many other countries, Denmark experienced economic deterioration in peripheral areas during the 1990es and onward, characterized by falling population growth rates and falling income. Therefore, in 2004, a tax exemption reform was established whereby citizens in selected peripheral municipalities were allowed a more generous tax reduction for transport to work. The intention was to stimulate more commuting, so that the selected municipalities would experience increased out-commuting, increased population growth, and increasing tax deductible income. Spatial interdependence is present as the effect on a treated municipality depends on whether neighboring municipalities are treated. Only by applying a spatial DID model do the theoretically predicted effects materialize.

1. Introduction

It is well known from social sciences that the specific circumstances in which a given policy is implemented can have important implications for the resulting outcome. Successful policies in one region or country may not have the same success when transferred to another country. A well-established econometric method to gauge if a policy (change) have the desired effect is the difference in difference approach, where change over time for the treated subjects is compared with the change for untreated subjects. An important assumption behind the method is that all subjects behave in the same way apart from the treatment effect and that potential outcome for one subject are unrelated to the treatment status of other subjects. If the assumptions are violated, estimation results become void. In a spatial context, locational differences interregional spillovers between regions may interfere with the DID model's assumptions. While behavioral differences between the subjects may be captured by inclusion of explanatory variables, an extension of the DiD model that enables us to cope with locational differences is very welcome. The spatial Difference-in-Difference approach recently

suggested by Delgado and Florax (2015) offers one way to include special spillover effects and thus improve estimations results.

The spatial DID model proposed by Delgado and Florax (2015) has been used by Chagas et al. (2016) to demonstrate the negative impact on respiratory diseases of sugarcane production in Brazil. The ability to distinguish between the internal effect in regions, where sugarcane production takes place and the spillover effect in other regions, a higher negative effect on respiratory diseases is found. Dube et al (2014) apply a spatial DiD model to estimate effects on housing prices of a change in public mass transit systems in Montreal, Canada. The spatial spillover stems from the assumption that the growth of housing prices is influenced by the rise in housing prices in the vicinity. Taking the indirect spillover effects into account gives a modest increase of the positive effect on housing prices of an improvement of public mass transit systems.

The case we use to demonstrate the advantages of Delgado and Florax (2015)'s extension is convenient because there is a clear economic dependence between the treated and the untreated regions where the impact of the treatment depends on whether the neighboring regions are treated. It furthermore turns out that the coefficient to treatment shifts from insignificance to significance and for one dependent variable shifts sign between the DiD model and the spatial DiD model. Only the spatial DID model gives the theoretically expected signs at a significant level.

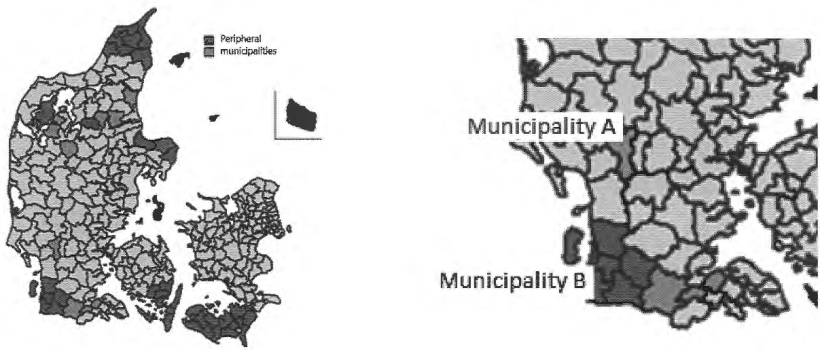


Figure 1. Peripheral municipalities 2004-2006

Denmark like most other countries experienced economic deterioration in peripheral areas during the 1990es and onward, characterized by falling population growth and falling income. Therefore, in 2004, a tax exemption reform was established in Denmark, so that citizens in peripheral municipalities were allowed a more generous tax reduction for transport to work (Skatteministeriet, 2009). In total 54 municipalities were appointed peripheral partly because of remote geographical location or because of a low municipal tax base. The selection method secures a reasonable degree of homogeneity among the

treated municipalities. Figure 1 shows the municipalities which were given formal status as being peripheral and thus eligible to the reform.

The intention of the reform was to stimulate more commuting, so that the peripheral municipalities would experience increased out-commuting, increased population growth, and an increasing tax base. By applying a recently suggested spatial Difference-in-Difference approach, the present study investigates as to whether such effects were realized.

Due to a data break caused by the structural reform of 2007, which merged the 275 Danish municipalities into 98 new, it is not possible to compare data before and after 2007. Thus, given that the tax exemption was introduced to the municipalities before this year, data for the new municipalities cannot be used, given that no observations prior to the introduction of the reform are available. However, data from the old municipalities applies, as they are made up of data from before the reform (2000-2003) and after the reform (2004-2006).

The case is interesting because the impact of the reform depends on whether the contiguous municipalities are subject to the reform. Thus, the impact can be expected to be higher in municipality A than in municipality B, see figure 1, because the latter is enclosed by other peripheral municipalities. The reason is that there is nothing to gain for a commuting worker by settling in municipality B compared to settling in one of the surrounding municipalities whereas this is the case for settling in municipality A. Moreover, with employment possibilities being comparatively good in the non-peripheral municipalities the commuting distance for the average commuting worker is lower in the peripheral municipalities surrounding B than in A. The assumption behind the DiD model that the potential outcome for one subject are unrelated to the treatment status of other subjects is clearly violated.

An outline of the paper is as follows. First, Section 2 presents the methods of the study, i.e. the traditional and spatial Difference-in-Difference (DiD) specifications. Next, Section 3 describes the data to be applied for the study, followed by s presentation of the estimated models for out-commuting, population growth and income in Section 4. Finally, Section 5 rounds off with concluding remarks.

2. Methods

The Difference-in-Difference DiD model has become standard in evaluation of program or treatment effects (e.g., Ashenfelter, 1978; Ashenfelter and Card, 1985; Rubin, 1974; Holland, 1986). The traditional DiD is most commonly specified for individuals, but may straightforwardly apply to regional units as follows. Assuming a sample of n municipalities observed over T years, where y_{it}represents an outcome that can potentially be affected by a reform, a standard DiD model can be written on matrix form as

$$y = \alpha_0 + X\alpha_1 + \alpha_2 D + \alpha_3 T + \alpha_4 D \circ T + \varepsilon$$

in which X is a matrix of potential covariates, ε a vector of mean-zero error terms which are uncorrelated with D and T, D a vector form time invariant indicator assuming the value 1 for the treated municipalities, T a vector form municipality invariant indicator assuming the value 1 for treatment years, and $D \circ T$ an interaction term vector, formed by space-wise multiplication (the Hadamard product), so that it serves as an indicator assuming the value 1 for treated municipalities in treatment years. Thus, while D_i represents the overall difference between treated and non-treated municipalities, and T_t represents the overall effect of the treatment years, the interaction term $D_i T_t$ represents the additional effect of the treatment for the treated municipalities in years of treatment and thus the treatment effect, also denoted the Average Treatment Effect (ATE).

However, given that the observational units are contiguous regions, there is a risk of spatial spillover of the reform effect. If such spillover is ignored, an omitted variable bias may occur and thus invalidate the estimated treatment effect. To account for such treatment spillover, one may re-specify the interaction term catching the treatment effect as $(I_{nT} + \rho W_s)D \circ T$, with $W_s = I_T \otimes W$, where ρ is a spillover parameter assumed to be numerically smaller than 1, and $\otimes$ denotes the Kronecker product. The term W_s is thus an nT by nT block-diagonal matrix, where the T diagonal blocks are formed by the n by n contiguity matrix W defined by letting $w_{ij} = 1$ if municipalities i and j are neighbors and 0 otherwise and subsequently row-standardized (Anselin, 1988). Thus, the traditional DiD model can be respecified as a spatial DiD reading as

$$y = \alpha_0 + X\alpha_1 + \alpha_2 D + \alpha_3 T + \alpha_4 (I_{nT} + \rho W_s) D \circ T + \varepsilon$$

or

$$y = \alpha_0 + X\alpha_1 + \alpha_2 D + \alpha_3 T + \alpha_4 D \circ T + \alpha_5 W_s D \circ T + \varepsilon$$

with the restriction $\alpha_5 = \rho\alpha_4$. For $\rho = 0$, α_5 also becomes 0, whereby the traditional non-spatial DiD occurs. As shown by Delgado and Florax (2015), the ATE now becomes $\alpha_4 + \alpha_5 \overline{WD}$, where $\overline{WD}$ is the average proportion of treated neighbors. Following the same authors, the ATE can be viewed as a sum of two terms: The average direct treatment effect ADTE given by α_4, and the average indirect treatment effect AITE given by $\alpha_5 \overline{WD}$.

3. Data

Data were observed for 268 Danish municipalities (excluding seven municipalities on the islands of Bornholm and Ærø, where data breaks were present due to early mergers) for the years 2000-2006 inclusive.

The variables included were percentage change in population size since previous year; income defined as the municipal tax deductible base (in 2000 prices); out-commuting rate, defined as number of commuters in percentage of number of employed persons in the municipality; an indicator for being a peripheral municipality; and an indicator for reform period (2004-2006). Furthermore, the interaction term Peripheral*Reform was formed, together with its spatial lag W*Peripheral*Reform. Data are summarized in Table 1.

Table 1. Description of data

Variable	Definition	Mean	Std
Out-commuting	Number of out-commuters per 100 employed persons	50,56	15,94
Population growth	Percentage growth in population size since last year	0,26	0,81
Tax base	Tax deductible income per inhabitant (1000 DKK; 2000 prices)	116,93	21,86
Time trend	1=2000, …, 7=2006	4,00	2,00
Peripheral	1=peripheral, 0=otherwise	0,18	0,38
Reform	1 if year=2004-06, 0 otherwise	0,43	0,50
Peripheral*Reform	1 if peripheral and reform period, 0 otherwise	0,08	0,26
W*Peripheral*Reform	Proportion of peripheral if peripheral and reform, 0 otherwise	0,07	0,22

4. Results

Initially, the non-spatial DiD models are reported in Table 2.

Table 2. Estimated non-spatial Difference-in-Difference (DiD) models

	Out-commuting			Population growth			Tax base		
Variable	**α**	**T**	**Sign.**	**α**	**T**	**Sign.**	**α**	**T**	**Sign.**
Intercept	49,78	47,86	***	0,32	6,35	***	113,49	81,67	***
Time trend	0,61	1,70	*	0,02	1,32		1,79	3,74	***
Peripheral	-8,99	-7,19	***	-0,60	-9,90	***	-16,34	-9,78	***
Reform	-0,34	-0,23		-0,05	-0,72		-2,04	-1,02	
Peripheral*Reform	0,73	0,38		-0,35	-3,76	***	0,11	0,04	

It is seen that peripheral municipalities generally perform poorer than non-peripheral in terms of population growth and municipal tax base. Peripheral municipalities have also

fewer out-commuters. Apparently, the reform did not have the wanted effects for the municipalities, neither in general or for the peripheral municipalities, and for population growth the treatment effect even seems to be significantly negative.

The estimated spatial DiD models for tax deductible income, population growth and out-commuting are provided in Table 3.

Table 3. Estimated spatial Difference-in-Difference (DiD) models

	Out-commuting			Population growth			Tax base		
Variable	**α**	**T**	**Sign**	**α**	**T**	**Sign**	**α**	**T**	**Sign**
Intercept	49,78	48,1	***	0,32	6,44	***	113,4	81,8	***
Time trend	0,61	1,71	*	0,02	1,34		1,79	3,74	***
Peripheral	-8,99	-7,24	***	-0,60	-10,04	***	-16,34	-9,80	***
Reform	0,57	0,38		0,01	0,15		-1,37	-0,68	
Peripheral*Reform	9,70	3,75	***	0,28	2,24	**	6,69	1,93	*
W*Peripheral*Refor	-14,45	-5,10	***	-1,02	-7,41	***	-10,61	-2,79	***

As before, it is seen that peripheral municipalities generally performed poorer than non-peripheral. Furthermore, it is confirmed that the reform did not on average have effects for the municipalities in general. However, when adjusting for the spatial spillover, it did raise the three measures in the peripheral areas after the reform. Also, it is seen that if a peripheral municipality is surrounded by a high proportion of other peripheral municipalities, then the outcome measures are significantly reduced. As explained in the previous section, there are good economic reasons to expect this. The contra-signed effects may explain why the treatment effect appeared insignificant in the non-spatial specifications; the treatment effects and their spatial spillovers apparently outperformed each other.

Table 4. Estimated Average Treatment (ATE) effects

ATE	**Out-commuting**	**Population Growth**	**Tax base**
ATE (traditional)	0,73	-0,35	0,11
ADTE (spatial model)	9,70	0,28	6,69
AITE (spatial model)	-1,01	-0,07	-0,74
ATE (spatial model)	8,69	0,21	5,95

Finally, the ATEs for the three outcomes are shown in Table 4 for the traditional as well as the non-traditional DiD. For the latter, the split into the non-spatial ADTE and the spatial AITE effects are also shown.

It is seen from Table 4 that the traditional ATE values are seriously misestimated. Opposed to these, the spatial DiD provides values which are reasonable as suggested by economic theory. Also, it is seen that the direct treatment effect (ADTE) is modified by presence of peripheral (= treated) neighbors.

5. Conclusion

We use a Danish case to demonstrate that a traditional non-spatial approach seriously may misestimate effects of an economic intervention in a case where there is spatial interdependence between the treated regions. The case concerns a generous tax reduction for transport to work for inhabitants in selected remote municipalities aiming to stimulate commuting, so that the selected municipalities would perform economically better. Spatial interdependence is at play because the effect in one municipality depends on whether neighboring municipalities also are selected. Only by applying a spatial DID model do theoretically predicted effects materialize.

References

Angrist, J., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. J. Amer. Statist. Assoc. 91, 444–472.

Anselin, L., 1988. Spatial Econometrics: Methods and Models. Kluwer Academic Publishers.

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. Rev. Econ. Stat. 60, 47–57.

Ashenfelter, O., Card, D., 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. Rev. Econ. Stat. 67, 648–660.

Chagas, A., Azzoni, C., Almeida, A., 2016. A spatial difference-in-differences analysis of the impact of sugarcane production on respiratory diseases. Regional Science and Urban Economics, 59, issue C, 24-36.

Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models and Applications. Pion, London.

Delgado MS, Florax RJGM. 2015. Difference-in-differences techniques for spatial data: Local autocorrelation and spatial interaction. Economics Letters 137, 123–126.

Dube, J.; Legros, D.; Theriault, M.; Des Rosiers, F., 2014. Transportation Research: Part B: Methodological. 64, 24-40.

Gerber, A.S., Green, D.P., 2012. Field Experiments: Design, Analysis, and Interpretation. W.W. Norton and Company.

Holland, P., 1986. Statistics and causal inference. J. Amer. Statist. Assoc. 81, 945–970.

Imbens, G.W., Wooldridge, J.M., 2009. Recent developments in the econometrics of program evaluation. J. Econ. Lit. 47, 5–86.

Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688–701.

Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Ann. Statist. 6, 34–58.

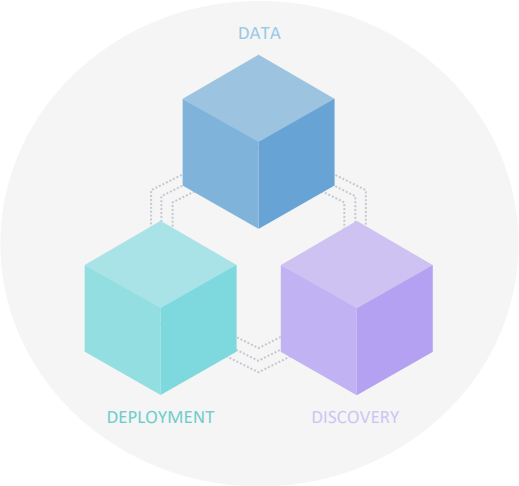
Rubin, D.B., 1990. Formal modes of statistical inference for causal effects. J. Statist. Plann. Inference 25, 279–292.

Skatteministeriet. 2009.

http://www.skm.dk/skattetal/statistik/tidsserieoversigter/befordringsfradraget-en-historisk-oversigt (viewed November 13, 2019).

What is SAS® Viya® for Learners?

SAS® Viya® for Learners delivers free access to advanced analytics software for teaching and learning. It is a suite of cloud-based software that supports the entire analytics life cycle – from data, to discovery, to deployment – and lets you code in SAS, Python or R.



Get Access

As an academic educator, you can apply for access by sending an email to: Academic.Nordic@sas.com