SYMPOSIUM I ANVENDT STATISTIK

2016

Copenhagen Business School
Institut for Finansiering og Økonomisk Institut

Danmarks Statistik
DST Survey

SYMPOSIUM I ANVENDT STATISTIK

25.-27. januar 2016

Redigeret af Peter Linde
på vegne af organisationskomiteen

Støttet af SAS Institute Inc.

CBS, Institut for Finansiering og Økonomisk Institut Danmarks Statistik

29. januar 2016: Revideret i forhold til oprindeligt offentliggjort.
Et indlæg var ikke kommet med i første udgave (Hedonic based price index)

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 Copenhagen Business School vært for symposiet, hvilket vi gerne vil takke for. Symposiet arrangeres af Danmarks Statistik og Institut for Finansiering og Økonomisk Institut på CBS og den faglige forening Symposium i Anvendt Statistik er ansvarlig for det faglige program og økonomien.

Denne publikation indeholder foredragene fra det 38. 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 978-87-501-2211-1

Organisationskomiteen for Symposium i Anvendt Statistik 2016

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

Peter Linde
DST Survey
Danmarks Statistik
Sejrøgade 11
2100 København Ø
pli@dst.dk

Anders Milhøj
Økonomisk Institut
Københavns Universitet
Studiestræde 6
1455 København K
Anders.Milhoj@econ.ku.dk

Esben Høg
Matematiske Fag
Aalborg Universitet
Fredrik Bajers Vej 7
9220 Aalborg Ø
esben@math.aau.dk

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

Anders Holm
Sociologisk Institut
København Universitet
Øster Farimagsgade 5
1014 København K
ah@soc.ku.dk

Helle M. Sommer
Afdeling for statistik og dataanalyse
DTU Compute
Danmarks Tekniske Universitet
2800 Lyngby
hems@dtu.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

Kaare Brandt Petersen
SAS Institute
Købmagergade 7-9
1050 København K
Kaare.Brandt@sdk.sas.com

Birthe Lykke Thomsen
Det Nationale Forskningscenter for Arbejdsmiljø
Lersø Parkallé 105
2100 København Ø
blt@arbejdsmiljoforskning.dk

Indholdsfortegnelse

Tidsrækker

Statistiske metoder og SAS

Samfundsanalyse 1

Samfundsanalyse 2

Markedsanalyse

Modeling Danish sperm donor limits

Claus Thorn Ekstrøm

Biostatistics, Department of Public Health, KU

Abstract

Danish sperm donors have a legal limit of 12 live children born from artificial insemination. This limit is based on similar numbers from other European and Western countries but no analysis based on Danish data has been presented. In this manuscript we will present and extend the Hajnal-Curie-Cohen model for modeling the expected number of incestuous marriages due to anonymous sperm donors and provide data relevant for Denmark and Danish sperm donations.

The result from the Danish data will be discussed and the risk of minimizing the genetic diversity in the Danish population due to the legal limits for the number of offspring that an anonymous donor can father will be investigated.

Introduction

The number of Danish children that a Danish sperm donor was allowed to father was 25 until October 2012 when the Danish sperm donation laws were changed. The laws were primarily changed due to an incident where a single donor was found to have fathered 43 children and passed on the genetic nerve disorder Neurofibromatosis Type 1 (NF1) or Von Recklinghausen's disease to at least five of the children conceived. Although the donor had fathered 43 children both within and outside Europe, the new limits were lowered (somewhat arbitrarily) to 12 to minimize the risk of segregating genetic mutations to a large number of offspring.

Most countries have restrictions on the number of children/pregnancies that an anonymous artificial insemination donor (AID) can father. These restrictions are set in place to reduce the risk of inbreeding and to minimize the risk that a sperm donor has an undiagnosed hertiable disease that is segregated to a large number of offspring. Children born from anonymous sperm donors, in particular, may result in an incestruous coupling and their children are more likely to become infected with recessive genetic diseases — in particular if the half-siblings are more likely to become a pair since they may be more likely to pair up due to assortative mating: age-wise, phenotypic-wise, and geography-wise the half-siblings look more like each other that a random person from the population.

Table 1 lists information about the sperm donation laws for countries similar to Denmark (obtained from Wikipedia on January 1st, 2016). The table shows that there is high variability in the number of children that a donor is allowed to father even for countries that are similar in culture and size to Denmark. The limits found in the countries have been set based on various historical, cultural, and mathematical evaluations, but there is really little concensus on how to set the limits.

Table 1: Table 1: Subset of information about sperm donor limits from Wikipedia

Country	Children.per.donor	Anonymity	Allowed.recipients
Belgium	6	varies	no data
Canada	25 per pop. of 800,000	no	no data
Denmark	12	varies	Everyone
France	5	yes	no data
Germany	15	no	Married heterosex. couples
Netherlands	25	no	Everyone
Norway	8	no	no data
Spain	6	yes	Everyone
Sweden	12 to 6 fam. (2 per fam.)	no	Married/cohabiting
Switzerland	8	no	Married heterosexual couples
United Kingdom	10 families worldwide	no	Everyone

In this manuscript I will present the Hajnal-Curie-Cohen (HCC) model for estimating the expected number of consanguineous matings that may occur. Relevant and recent Danish data will be used to estimate the expected time to an incestuous marriage occurs. Furthermore, the HCC model will be extended to be able to compute relevant confidence and prediction limits for the probability that an incestreous coupling occurs. Finally, a few comments about the changes in inbreeding coeffiecient are discussed in the Discussion.

Methods

The Hajnal-Curie-Cohen (HCC) model (see Hajnal 1960; Curie-Cohen 1980) is widely used to evaluate the average number of potential unwittingly consanguineous half-sibling matings among the offspring of an anonymous artificial insemination sperm donor. The popularity of the HCC model is partly based on the fact that it has in many situations been possible to obtain the necessary input data or guesstimates in order to estimate the number of AID-based half-sib matings per year.

Let Y be the number of half-sibling matings for an anonymous single sperm donor and let S be the number of new sperm donors per year in a country. The HCC model proposes that

$$E(Y) = S\bar{m}\pi,$$

where $\bar{m}$ is the expected number of potential half-sib matings per donor, and π is the probability of a mating between a random male and female half-sib, which may depend on assortative mating. Note that in the HCC model we have implicitly assumed that probability of mating between two offspring of the same donor is so low that we can neglect the chance that two or more such matings occur. S is a number that often can be easily obtained from hospital records or clinic and $\bar{m}$ is easily computed as will be shown below. The major problem in the HCC model is to provide a good estimate for π.

Incestuous half-sibling matings occur either when two AID children become a couple or when an AID offspring pairs up with one of the donor's natural children. If the donor has on average of $\bar{f}$ natural children and $k = b + g$ AID children (with b boys and g girls) then on average (assuming that boys and girls are equally prevalent among both the natural and AID children) we have that the expected number of potential pairs becomes

$$\begin{aligned}\bar{m} &= E(b \cdot (k-b)) + E(b) \cdot \bar{f}/2 + E(g) \cdot \bar{f}/2, \\ &= \frac{V(k) + \bar{k}^2 - \bar{k}}{4} + \frac{\bar{k}\bar{f}}{2}\end{aligned}$$

where the first element in the formula is due to AID offspring pairings and the last two elements are due to pairings between an AID child and a natural child. Furthermore, we have assumed that there is an (unspecified) distribution of the number of AID children that an AID donor produces, k. The second equality sign follows from properties of the binomial distribution.

If every donor produces exactly k AID children — which in reality is the situation that a government should consider — then the formula for $\bar{m}$ reduces to

$$\bar{m} = \frac{k(k-1)}{4} + \frac{k \cdot f}{2}.$$

Estimating the probability of mating between half-sibs

As mentioned above, the probability that two half-siblings mate, π, may be difficult to estimate since it is virtually impossible to find data on half-sib matings because they are illegal in most countries. Instead, the HCC model base the estimate of π on loose assumptions about assortative mating derived from the half-sib pair's age difference, geographic location, and phenotypic similarity.

Consider a random pair of paternal half-sibs. If the age difference between the pair (male-female) is r then we have that the probability that that specific pair mates is

$$\begin{aligned}\pi &= P(\text{pair mates}|\text{half-sibs}) \\ &= P(\text{pair mates} \wedge \text{mate with age diff } r|\text{half-sibs}) \\ &= P(\text{pair mates}|\text{half-sibs, mate with age diff } r)P(\text{mate with age diff } r|\text{half-sibs}).\end{aligned}$$

Hajnal (1960) and Curie-Cohen (1980) show that the probability that two half-siblings mate, π, can be simplified to

$$\pi = \frac{l\bar{d}CQ}{A/2},$$

where l is the probability that a newborn will reproduce, $A/2$ is the size of the available population of potential mates for the male (the total population of eligible singles per year is

A), and Q is the proportion of marriages/couplings that happen between people from the same region. If there are multiple (separate) regions in the country then the HCC model should sum over all regions, i.e., SQ/A becomes

$$\sum_{\text{region } j} \frac{Q_j S_j}{A_j}$$

$\bar{d}$ is the (average) probability that a pair will end up matings depending on their age difference r averaged over the distribution of the age difference between the AID children. In particular, $\bar{d} = \sum \phi(r)\psi(r)$ where $\phi(r)$ is the proportion of matings where the woman was exactly r years younger than the male in the population, while $\psi(r)$ is the age difference distribution for the half-siblings. Finally, Cavalli-Sforza, Kimura, and Barrai (1966) show that for I (independent) phenotypic traits that influence mating patterns (ie., result in assortative mating) we get

$$C = \prod_{i=1}^{I} \frac{1}{1 - \xi_i \rho_i},$$

where ρ_i is the phenotypic correlation between mates/spouses for trait i while ξ_i is an (indirect) measure of the heritability of the trait. In the paper by Curie-Cohen (1980) some loose arguments are used to obtain a value for C of 2.

To summarize, the HCC model can be written as

$$\begin{aligned} E(Y) &= S\bar{m}\frac{l\bar{d}Q}{A/2}\prod_{i=1}^{I}\frac{1}{1-\xi_i\rho_i} \\ &= 2\bar{m}l\bar{d}\prod_{i=1}^{I}\frac{1}{1-\xi_i\rho_i}\sum_{\text{region } j}\frac{Q_j S_j}{A_j} \end{aligned}$$

Extending the HCC model

The HCC model estimates the *expected* number of cosanguinuous matings for half-sib pair, but if we are worried about the risk of cosanguinuous matings then we also need to * estimate the variation among donors to provide confidence intervals, and to * take into account any influence that the age patterns may have on the risk of assortative matings among the half-sibs. Finally, the phenotypic influence on the mating pattern should be updated so that it is not just based on ear-length, body stature, and IQ (which is the basis found in Curie-Cohen (1980)). These extensions are achieved by using distributions for the relevant terms in for HCC model in order to remove the expectation, and provide a proper model for Y.

Danish data

To compute the limits for Denmark we need to specify the following values for the HCC model

1. S the number of new donors per year. According to *Cryos International* this is fairly constant around 40. Since there are two major providers of anonymous sperm donors in Denmark (one centered in Copenhagen and one in Aarhus) we can consider this to be 80 (40 for each region).
2. l is the probability that a newborn baby will reproduce. According to *Statistisk Årbog* this value can be set at 0.884.
3. $\bar{d}$ models the probability of mating due to the age difference among the half-sibs. For this parameter we can use information from *Cryos International* to see how the age distribution of the donors compare to the age distribution of fathers in the general population. More information on this below since we intend to use the actual age distribution.
4. According to *Danmarks Statistik* there were a total of 828504 singles below the age of 50 in all of Denmark on January 1st, 2015. The parameter A_j is the number of eligible singles that an offspring of a single donor might marry (both men and women), and if we assume that there are 5 regions in Denmark then this gives a total of 165700 "fish in the water" for each region. Half of them are assumed to be male and the other half female, although the age distribution is slightly different between the sexes.
5. Q is the proportion of coupling that are made with people from the same region as the person himself/herself. We do not know the proportion of matings that occur with people from outside each region, but a general overall guesstimate for the emigration for each region would be $Q = 6/7$.
6. The coefficient that depends on assortative mating due to similarity, C (and the correlation and heritability), will be discussed below.

We vary the number of AID children, k and assume that all donors get exactly k AID children. This will result in an overestimate of the actual risk of inbreeding due to AID since some donors will in reality result in fewer than k AID children.

Curie-Cohen (1980), Silventoinen et al. (2003), Zietsch et al. (2011) and Keller et al. (2013) consider correlations and heritabilities for various genetic-related phenotypic traits that might have influence on the choice of partners (see the table).

Trait	Correlation	Heritability	C
Height	0.20	0.75	1.0524931
IQ	0.35	0.57	1.0790486
Ear length	0.40	0.75	1.1481056
Stature	0.35	0.88	1.1530469
IQ (study 2)	0.33	0.69	1.0954415

The overall level of the C component for the traits listed above seems somewhat stable, but while there are several papers that show that a specific genetic-based trait influences mating, the effective number of (independent) traits is unknown. Curie-Cohen (1980) suggest to use the value 2 (or possibly 4 to err on the side of safety). Here we will assume a prior distribution on the C such that we first draw the effective number of traits from a uniform distribution on $\{4, 5, 6\}$ and then draw each of the traits from a log-normal distribution such that

$$C = \exp\left(\sum_{\text{traits t}} Z_t\right), \text{ with } Z_t \sim \log N(0.13, 0.02^2)$$

This results in an average value of C around 1.9.

Donor data summary

In order to provide realistic information on the distribution of the actual ages of anonymous sperm donors and the number of Danish offspring that they father. The distribution of the Danish donors (from the years 2001–2010) are shown. We have remove donors from 2011-2015 to make sure that they are all have had time to produce up to 12 offspring. For donors before 2012 we have restricted to the offspring to the first 12 if there were more offspring than that.

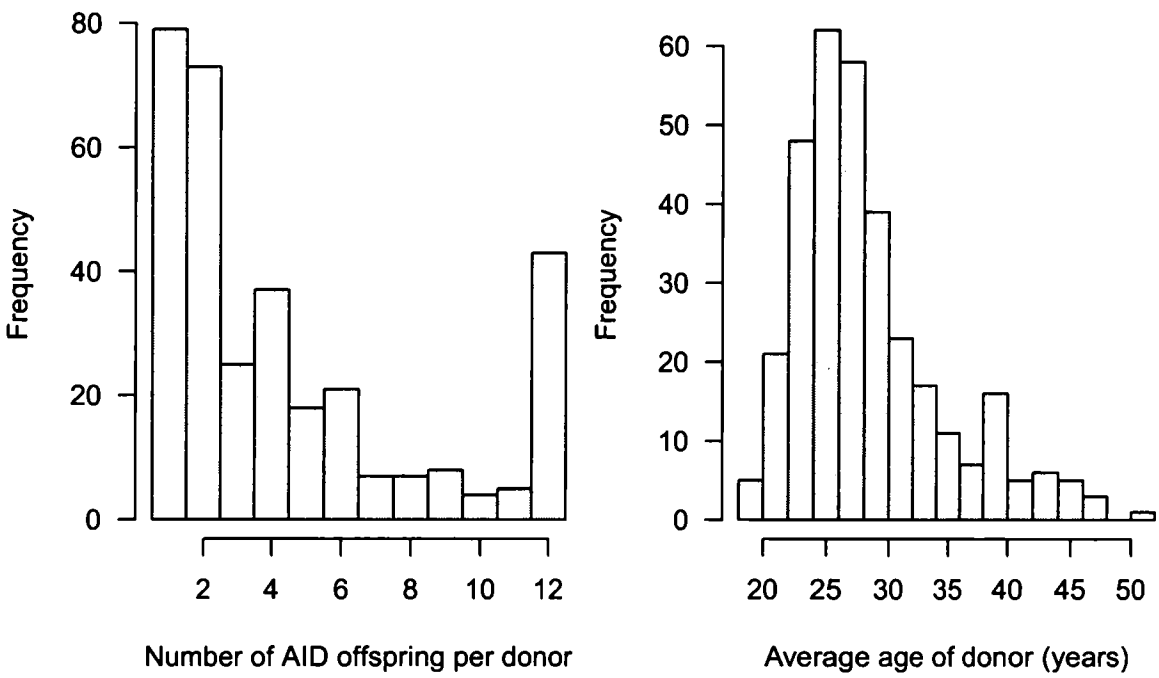


Figure 1: Distribution of number of actual AID offspring per donor (left), and average age of donor at time of pregnancy (right)

When modeling the age distribution of the donors it is worth seeing how the span over calendar years depends on the number of AID children the sponsor fathers. A heatmap

showing virtually no substantial association between the two is shown below. The only noticeable deviation is for donors with 12 offspring since those 12 offspring are typically produced over several years.

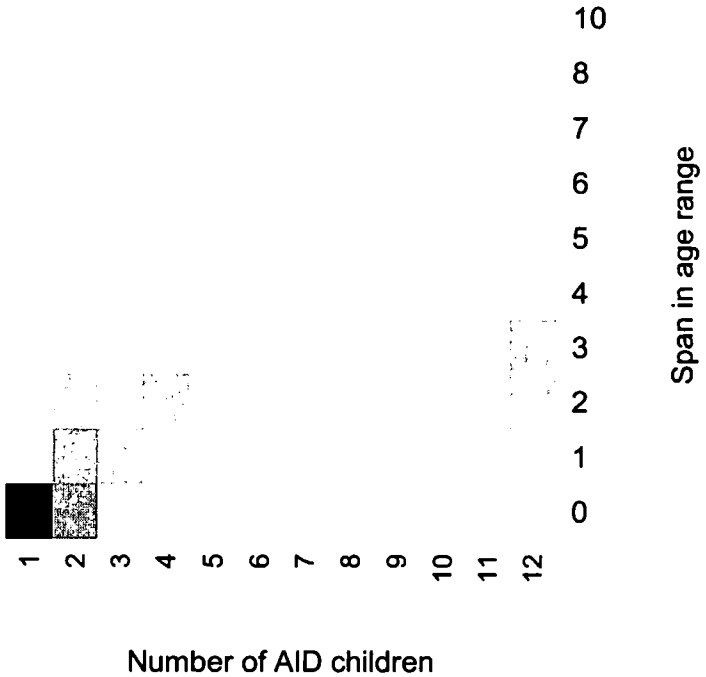


Figure 2: Heatmap of number of AID children and span in period over which the children were born.

The Danish population

We have obtained the data from Nordfalk, Hvidtfeldt, and Keiding (2015) on the Danish age patterns among the parents at the time they give birth. The distribution about the age difference can be seen in the histogram below. The Figure also shows the age densities in the population for fathers getting their first, second, .. etc child.

The average difference in ages for parents in the Danish population is 2.77 years.

Results

Here we compare the results from the HCC model using the Danish data to the results from the HCC model under a "worst case scenario":

- All donors get the maximum number of donor children.
- The gender distribution for children of a donor is exactly even.

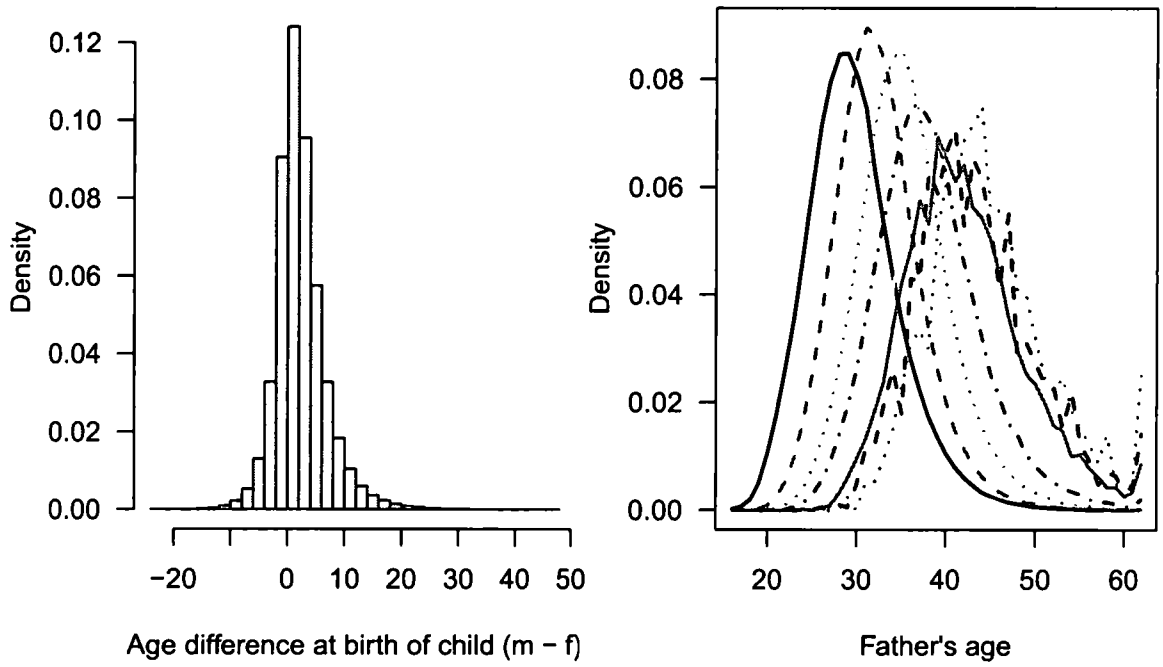


Figure 3: Distribution of age differences (left) and father's age for parities 1-8 (right) in the Danish population

- All children of the same donor have the same age such that all males are exactly 2.7 years older than females.

For each situation we compute the expected number of inbred matings per year and invert the result to get the expected number of years, ie., the recurrence time before an inbreeding occurs.

The results (based on the standard HCC model for Danish data) can be see in the figure below. For the current Danish legal limits the average time until an inbreeding occurs is around 180 years (based on the HCC model). For the worst case situations, ie., the HCC model with no gender variation and no age variation the recurrence time is somewhat smaller but still far from being a problem for the Danish society in practice.

The final figure shows the results from the extended HCC model. The wide area is the prediction interval and the darker area is the pointwise confidence interval for the expected number years before an consanguinuous mating occurs.

There are two important messages to get from the extended HCC model analysis of the Danish data. First, the extended HCC model shows that the standard HCC model substantially overestimates the risk that an inbred mating occurs. Despite the AID children being born within a very small time span, the extended HCC model still yields longer expected time to an inbred pairing. This is reassuring because it means that the potential increase that might have occurred because half-sibs are similar really did not have a major impact. Also, we see

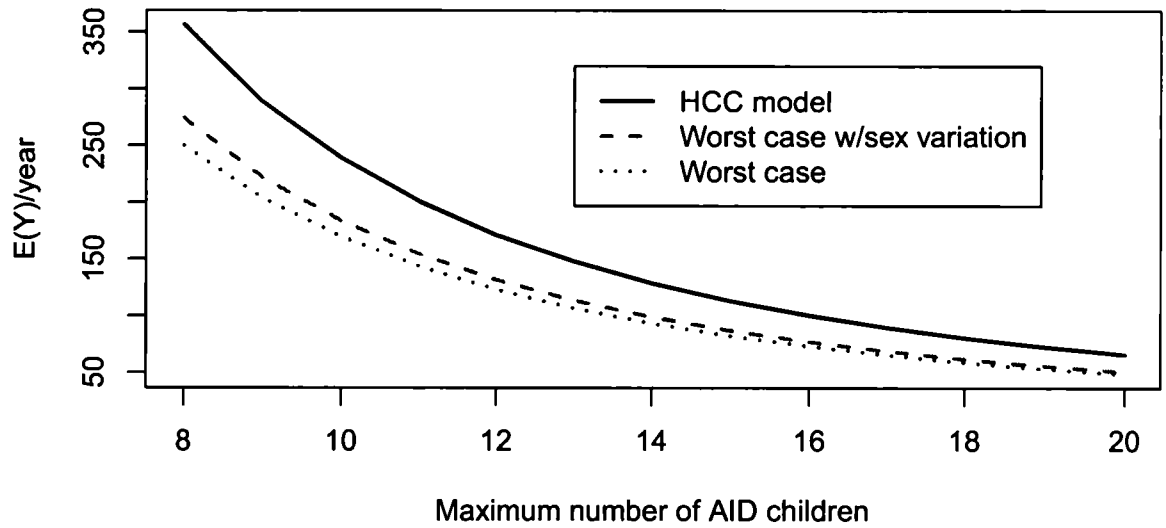


Figure 4: Danish estimates from the HCC model

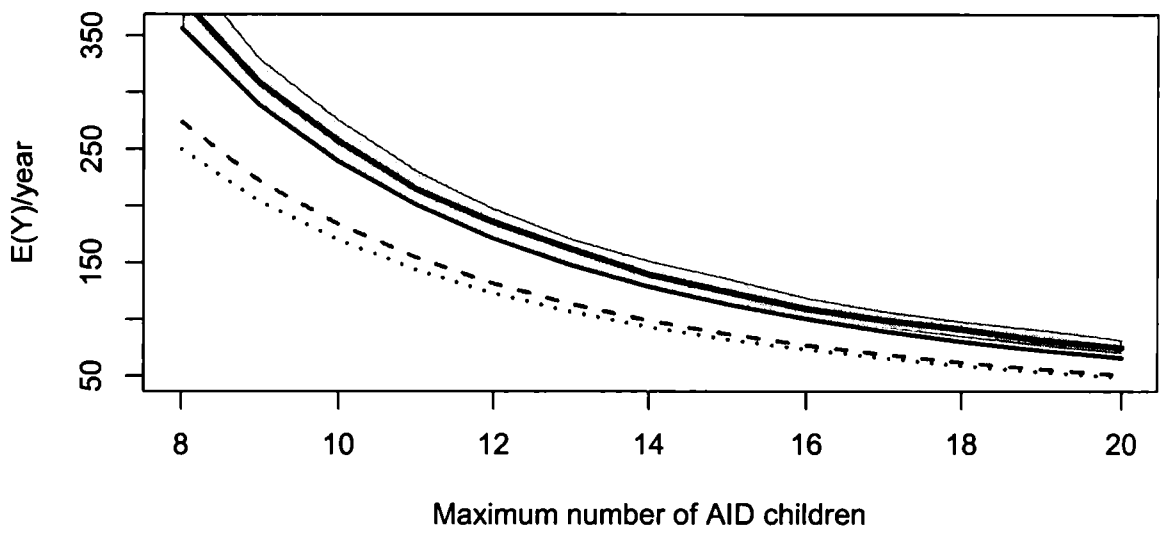


Figure 5: Danish estimates, CI, and PI from the extended HCC model

that the expected recurrence time results are fairly stable but there may be somewhat of an individual variation that result in large prediction intervals.

Discussion

The change in legal limits from 25 off spring for a Danish sperm donor to 12 offspring has increased the expected time on an inbred mating between half-sibs roughly by a factor 4. Using realistic Danish data we have an expected time until a mating that is around 250 years which must be said to pose next to no inbreeding problems in Denmark.

If we were to consider the impact of imbreeding in the Danish population we could consider the looking at the average coefficient of inbreeding, F. The coefficient of inbreeding measures the expected percentage of homozygosity arising from a given system of breeding, and in our case this is AID offspring pairings. A general consequence of inbreeding is that homozygotes increases and that the number of heterozygotes decrease.

Bibliography

Cavalli-Sforza, L. L., M. Kimura, and I. Barrai. 1966. "The Probability of Consanguineous Marriages." *Genetics* 54 (1): 37–60.

Curie-Cohen, Martin. 1980. "The Frequency of Consanguineous Matings Due to Multiple Use of Donors in Artificial Insemination." *American Journal of Human Genetics* 32: 589–600.

Hajnal, John. 1960. "Artificial Insemination and the Frequency of Incestuous Marriages." *Journal of the Royal Statistical Society, Series A* 123 (2): 182–94.

Keller, Matthew C., Christine E. Garver-Apgar, Margaret J. Wright, Nicholas G. Martin, Robin P. Corley, Michael C. Stallings, John K. Hewitt, and Brendan P. Zietsch. 2013. "The Genetic Correlation Between Height and IQ: Shared Genes or Assortative Mating?" *PLoS Genetics* 9 (3): e1003451.

Nordfalk, Francisca, Ulla A. Hvidtfeldt, and Niels Keiding. 2015. "TFR for Males in Denmark – Calculation and Tempo-Correction." *Demographic Research* 32: 1421–34.

Silventoinen, Karri, Jaakko Kaprio, Eero Lahelma, Richard J. Viken, and Richard J. Rose. 2003. "Assortative Mating by Body Height and BMI: Finnish Twins and Their Spouses." *American Journal of Human Biology* 15 (5). Wiley Subscription Services, Inc., A Wiley Company: 620–27.

Zietsch, Brendan P., Karin J. H. Verweij, Andrew C. Heath, and Nicholas G. Martin. 2011. "Variation in Human Mate Choice: Simultaneously Investigating Heritability, Parental Influence, Sexual Imprinting, and Assortative Mating." *Am Nat* 177 (5): 605–16.

Trends in death rates from diabetes in Denmark, 1994-2028

Gustav N Kristensen
Stockholm School of Economics in Riga
kristensengn@gmail.com

Maja Sparre-Sørensen
University Hospital Herlev
majasparre@gmail.com

Abstract
The rapid increase in diabetes has worried the health sector.
Age is the dominating factor in diabetes, but malnutrition played a significant role in the years 1999-2007 in all age groups above the age of 45.
Protective and detrimental cohorts plays an important role in the development of the death rate from diabetes over time. However, the trend showed the most peculiar development, which attracted the attention of the authors of this article. Although the time trend shows a slight increase, that development is outweighed by the protective cohort effect and the combined result is a long range decrease (predominately for women) in the death rate from diabetes.

Keywords: diabetes, trend, cohorts, malnutrition, death rates.

Introduction

Diabetes is a merciless disease that invalidate and kills people in a huge number. Therefore concern for its development is justified[1]
The Danish malnutrition period described in [1] led to several studies of the death rate from diabetes provoked by malnutrition [2-5].
A number of studies of the impact of the Dutch famine where the effect of the starvation in the period October 1944 to May 1945 on children exposed to prenatal famine have an increased risk of developing a number of serious diseases over lifetime [6-8]. This tragic event has increased the interest for Age-Period-Cohort studies [9-16].

Lately the huge (worldwide) increase in diabetes has called the attention of researchers in public health. In 1995 the number of patients suffering from diabetes worldwide was approximately 135 million, in 2025 the number is estimated to more than double to 300 million. Among the Danes 80-85% of all patients suffer from lifestyle induced type 2 diabetes, which increases with age [17].
This raise an interest in studying the trend-development in the death rate from diabetes on Danish data.

The purpose of this study is to model the trend development in the death rate from diabetes in the period 1994-2028, with ex ante forecast of death rate from diabetes for the period 2014-2028.

Data

The data material is coincide with the data set from [18] we likewise applied in [1-5].
Death rate is the number of death from a certain cause per 100 000 persons in a considered group. (E.g. the number of 80-84 year old women death from diabetes per 100 000 women in the age group 80-84.)
The Danish data on the *death rate* from malnutrition and diabetes are taken from: The State Serum Institute: Malnutrition, B-040 and Diabetes (Sukkersyge) B-039.
There is no distinction between different types of diabetes.
The variables applied for the model of the death rate from diabetes are:

Diab	death rate from diabetes
D2000m/D2001w	dummy for the years with a malnutrition choc
T	trend - the period (or year), 1977 = 1
Age	age at death
CohYear	diagonal dummy system indicating year of birth
B	coefficient vector – the estimated Beta-string
Gender (index)	*m* – men *w* – women

[1] Ekstra Bladet 8. sep. 2015 writes on its front page: "Nu springer diabetes bomben". (Now the diabetes bomb explodes), and publish choc data for diabetes until 2040 based on a linear trend.

The "trend" describe a general (global) development based on none or only very vague theory on its origin like "the technical progress" or "the economic growth".

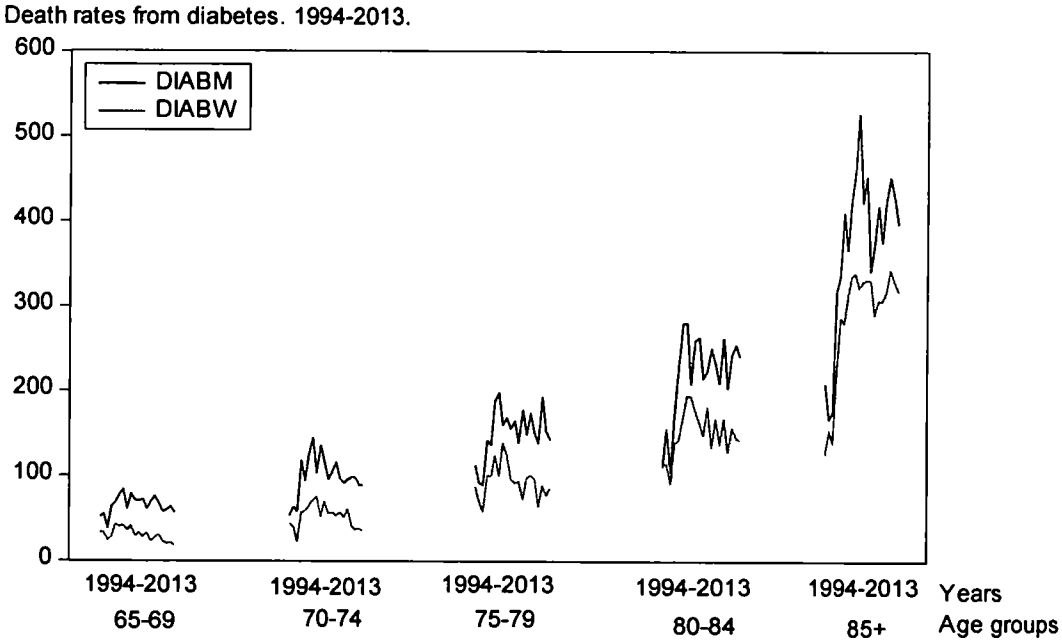


Figure 1. Death rates from diabetes 1994-2013.

Figure 1 shows the death rate from diabetes 1994-2013 for age groups from 65.
The 9 age groups are shown in Table 1:

Table 1. Age groups and years of birth.

Age	Born	Data years
85+	1909-1928	1994-2013
80-84	1914-1933	1994-2013
75-79	1919-1938	1994-2013
70-74	1924-1943	1994-2013
65-69	1929-1948	1994-2013
60-64	1934-1953	1994-2013
55-59	1939-1958	1994-2013
50-54	1944-1963	1994-2013
45-49	1949-1968	1994-2013

Ex-post and ex-ante forecasts

The estimated Beta-string (see below equation 2) gives information about past and future protective and detrimental cohort effects associated with diabetes. Those effects can be applied in prognosis for the death rates from diabetes.

Table 2. Age groups and years of forecasts.

Age	Ex-post	Actual	Ex-ante
85+	1909	1994-2013	2053
80-84	1914	1994-2013	2048
75-79	1919	1994-2013	2043
70-74	1924	1994-2013	2038
65-69	1929	1994-2013	2033
60-64	1934	1994-2013	2028
55-59	1939	1994-2013	2023
50-54	1944	1994-2013	2018
45-49	1949	1994-2013	2013

Models

The model gives a description of the development in the death rate from diabetes distributed on men and women. Age is the dominant explaining variable, then comes a trend, a variable for the Danish malnutrition period, a dummy for a malnutrition choc, and finally in equation (2) the cohort effect all creating an Age-Period-Cohort model.

$$\text{Log}(Diab) = \alpha_1 + \alpha_2 Age + \alpha_3 Age^2 + \alpha_4 Sqr(T) + \alpha_5 Sqr(Dmal) + \alpha_6 D2000 + \text{er} \quad (1)$$

The assumption behind the cohort dummies (see Kristensen [14]) is that each cohort throughout life remain in the same health group, e.g., the 85-year-old in 1983 is in the same health group as the 80-year-old in 1978. For estimation the model was formed as

$$\text{Log}(Diab) = \alpha_1 + \alpha_2 Age + \alpha_3 Age^2 + \alpha_4 Sqr(T) + \alpha_5 Sqr(Dmal) + \alpha_6 D2000 + \beta_{10} Coh1910 + \beta_{11} Coh1911 + \ldots\ldots + \beta_{67} Coh1967 \quad (2)$$

The coefficients in (1) and (2) were calculated by Weighted Least Square using "Age" as weight. The model was estimated for men and women. For women α_3 and α_4 took the value 0. Among the dummies Coh1909 was chosen as reference cohort and consequently omitted in the estimation (or more technically expressed included in the constant element α_1)

As an alternative we can see the cohort effects as "the residual" and make a twostep estimation by a separate estimation of the beta coefficients. Thereby we can have an estimate for β_9 and β_{68}.

$$\text{er} = \beta_9 Coh1909 + \beta_{10} Coh1910 + \beta_{11} Coh1911 + \ldots\ldots + \beta_{68} Coh1968 \quad (2a)$$

We here do not have to drop a dummy. We here disregard the remaining residual.

The partial cohort effect

The Beta-string or the estimated beta-coefficients are shown in Figure 2. The Second World War and the time after reduce the beta-coefficients. For women 1939-1949, for men 1939-1951.

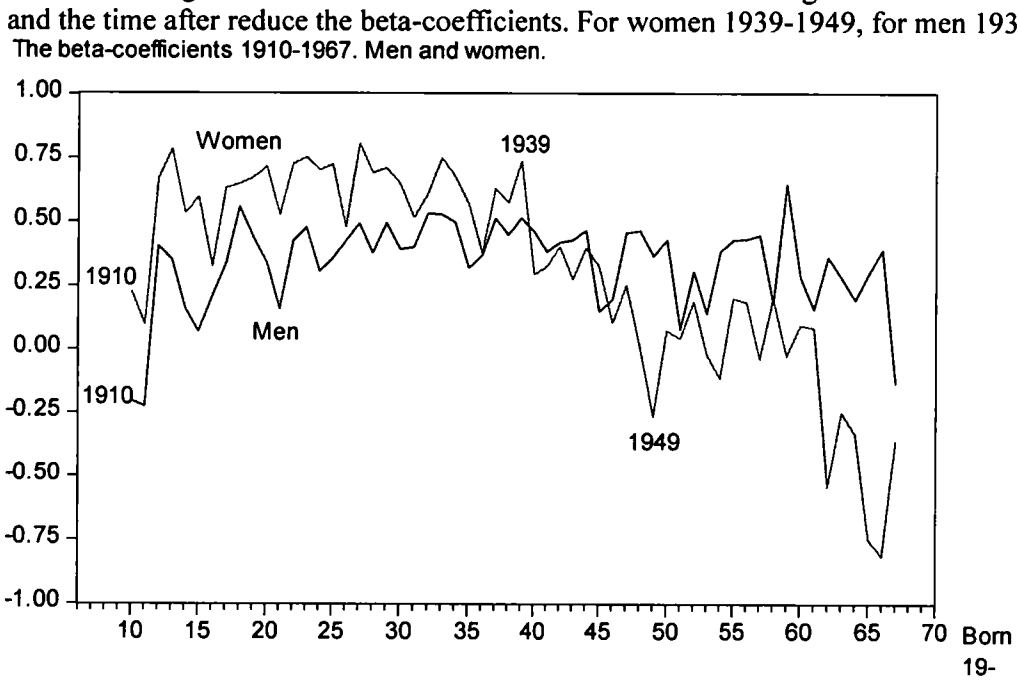


Figure 2. The estimated beta-coefficients for men and women.

As shown in [5 and 14] we can create a new and handier variable from the beta-coefficients in the estimated equation (2),
Sections of the beta-coefficients can be seen as time series explaining the death rate. Sections related to a certain age groups is indicated by a foot-number

The last person on 45 years was born in 1968. Last cohort is born in 1968.

It is seen from Table 2 that the Beta-string applied as explaining variable can give data to forecast up to 2053.

Forecast up to 2028

The estimated model can using the estimated cohort coefficients forecast until 2053, however, with increasing risk that the assumption on the time trend do not hold. We therefore stops by 2028.

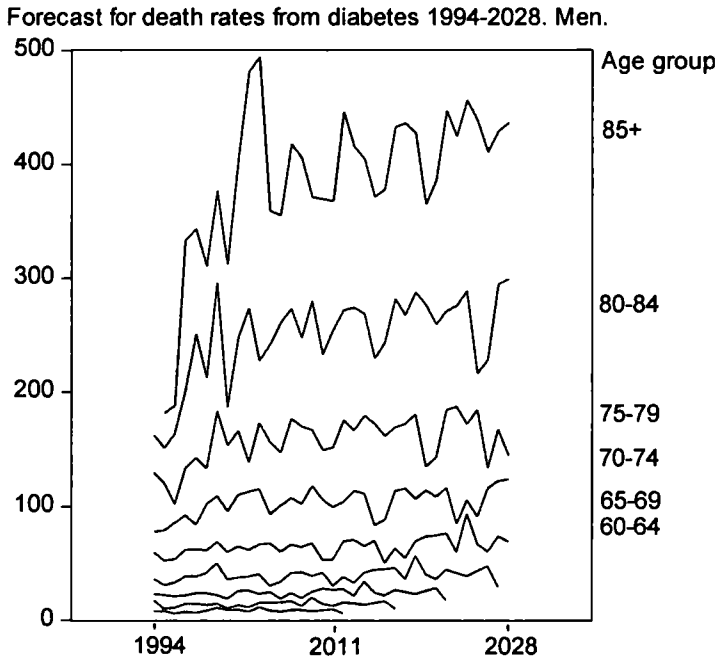


Figure 3. Forecast for death rates from diabetes 1994-2028. Men.

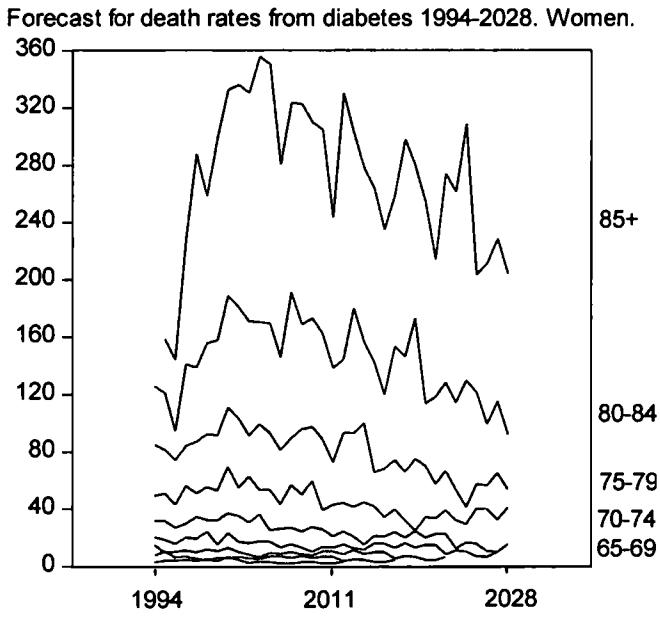


Figure 4. Forecast for death rates from diabetes 1994-2028. Women.

Discussion

Underweight and weight loss increases mortality in both type 1 and type 2 diabetes patients but in two different ways.
We were in our research not able to distinguish between type1 and type 2 related deaths; we therefore cannot say if there will be a significant difference in the death rate between to two diseases. No reservation for invalidation caused by diabetes was made.

I this studies we only consider diabetes related deaths. The lifelong effect of diabetes on morbidity is not examined
We cannot not say for certain why World War 2 effected men and women differently. It seems that women for some reason are more robust than men.

The model assume no future malnutrition period as well as no future malnutrition choc.
The trend included as Sqr(T) is ever increasing, but will most likely change and become decreasing due to medical progress.

Conclusion

For death rates from diabetes no long range increase is seen in the Danish data.

References.

[1] Sparre-Sørensen M, and Kristensen GN (2015). Underernæringen i Danmark. 1999 – 2007. Hvordan tæller man de døde? *Symposium i anvendt statistik. Danmarks Tekniske Højskole*, 2015; 200-211.

[2] Sparre-Sørensen M, Kristensen GN (2015). Alzheimer's disease in the Danish Malnutrition period, 1999 – 2007. *Journal of Alzheimer's Disease*. **48** (4), 979-985.

[3] Sparre-Sørensen M, Kristensen GN (2015). Schizophrenia and the Danish malnutrition period, 1999 – 2007. *International Journal of Health Research and Innovation,* forthcoming.

[4] Sparre-Sørensen M, and Kristensen GN (2015). Malnutrition related deaths. (Paper in progress).

[5] Kristensen GN, Sparre-Sørensen M. (2015) Death from Stroke during the Danish malnutrition period 1999-2007. *Journal of Statistical and Econometric Methods*, **4** (2): 127-154.

[6] Hoek et al. (1998). The Dutch famine and schizophrenia spectrum. *Soc Psychiatry Psychiatr Epidemiol*. **33**(8): 373-9

[7] de Rooij SR (2013). Blunted cardiovascular and cortisol reactivity to acute psychological stress: a summary of results from the Dutch famine birth cohort study. *International Journal of Psychophysiology*, **90**; 21-27

[8] de Rooij et al. (2013). The developmental origins of ageing: study protocol for the Dutch famine birth cohort study on ageing. *BMJ Open 3: e003167.*

[9] Glenn ND (1976). Cohort Analysts' Futile Quest: Statistical Attempts to Separate Age, Period and Cohort Effects. *American Sociological Review*, **41** (5): 900-904.

[10] Osmond C, and Gardner MJ (1982). Age, Period and Cohort Models applied to Cancer Mortality Rates. *Statistics in Medicine*, **1**: 245-259.

[11] Clayton D, Schifflers E. (1987a). Models for temporal variation in cancer rates. I. Age-period and age-cohort models. *Statistics in Medicine,* **6**: 449-467.

[12] Clayton D, Schifflers E. (1987b). Models for temporal variation in cancer rates. II. The age-period-cohort model. *Statistics in Medicine,* **6**: 469-481.

[13] Holford TR (1991). Understanding the effects of age, period, and cohort on incidence and mortality. *Annu. Rev. Publ. Health* **12**: 425-457.

[14] Kristensen GN (2013). Cohort Coefficients. Describing the secular development in protective and detrimental cohort effects associated with apoplexy. *Journal of Statistical and Econometric Methods.* **2**(4): 119-127.

[15] Kristensen GN (2014a). Testing 'Clemmensen's hook' in the death rate from breast cancer. *Journal of Statistical and Econometric Methods.* **3**(2): (2014), 15-30

[16] Kristensen GN (2014b). The Holford puzzle. The cohort effects in death rates from lung cancer. *International Journal of Health Research and Innovation.* **2**(1): 1-14.

[17] The national institute for public health, Folkesundheds rapporten 2007, www-si-folkesundhed.dk/upload/kap_11_diabetes.pdf

[18] Statens Serum Institut (2014). *Database.*

Point-of-care testing of HbA1c in diabetes care and preventable hospital admissions

Troels Kristensen[a,c] & Kim Rose-Olsen[b,c]

[a] COHERE, Institute of Public Health, Faculty of Health Sciences
[b]COHERE, Department of Business and Economics, Faculty of Social Science
[c]Research Unit of General Practice, Faculty of Health Science
University of Southern Denmark,
[a,c]J.B. Winsløw vej 9 5000 Odense C & [b]Campus vej 55 5230 Odense M.

Work in progress, do not refer to or cite without permission from the authors.
This version: December 15, 2015

Background: Point-of-care testing (POCT) of HbA1c may result in improved diabetic control, better patient outcomes and enhanced clinical efficiency with fewer patient visits and subsequent reductions in hospitalizations and costs. In 2008, the Danish regulators agreed to create a new tariff for the remuneration of POCT of HbA1c in primary care.

Aim: The aim of this study is to assess whether there is an association between the use of POCT of HbA1c and preventable hospital admissions among diabetes patients in general practice.

Method: We apply logistic regression analyses to examine whether there is a link between preventable hospital admissions and POCT of HbA1c in general practice. Preventable hospital admissions were assessed through the ambulatory care sensitive conditions (ACSCs) classification of hospital admissions. We include independent variables such as gender, age, ethnicity, socioeconomic covariates, municipality classifications and case mix measure in terms of the charlson index and costs of care in primary care and secondary care.

Results: There was a significant link between POCT of HbA1c among diabetes patients in general practice and an ACSC-measure of preventable out- and inpatient hospital treatment after adjusting for individual-level patient characteristics. However, the link was not significant when the measure of hospital treatment for ACSCs was restricted to inpatient treatment.

Conclusion: Our preliminary results indicate that more POCT of HbA1c may result in fewer preventable hospital treatments for ACSCs among diabetes patients in general practice.

Introduction:

Preventable hospital admissions are those for which hospitalization is thought avoidable if preventive care and early disease management are applied in the ambulatory care setting such as general practice (Ansari et al. 2002). This means admissions to a hospital for certain acute illnesses or

worsening chronic conditions (e.g., diabetes) that might not have required hospitalization had these conditions been managed successfully by primary care providers in outpatient settings (Moy et al. 2013). In health care systems with high hospital admission rates for diabetes, medical care may be suboptimal because there is likelihood of hospitalization for mild metabolic problems (Connell 1985). Potential interventions may be primary prevention, case detection, treatment and monitoring as well as Workforce development programs for GPs to improve clinical practices. The Danish authorities have stated that citizens with diabetes are hospitalized more than in other countries like Sweden and Norway. The latter has also highlighted that there are large variations in hospitalization rates across Danish Regions and municipalities and a need for earlier diagnosis and treatment of chronic conditions. This development may raise concerns among policy-makers and health-planners. Several measures such as the Ambulatory Care Sensitive condition System (ACSC) and potentially preventable hospitalizations (PPH) have been developed to measure preventable hospital admissions (Gao et al. 2014; Moy et al. 2013). These population based measures may identify health care problems, help focus interventions to improve diabetes care and improve the efficiency of health systems by promoting ambulatory care as well as population health (Arrieta and Garcia-Prado 2015). Thus, measures of preventable hospital admissions may be used as an indicator of primary care access, assist primary care providers to identify high-risk patients for early intervention and to predict ACSC hospitalizations for diabetes (Gao et al. 2014).

Point of Care testing (POCT) of HbA1c of diabetes patients rather than laboratory testing may be one way to try to avoid hospital admission. The advantage is that near-immediate test results allow patients and doctors to evaluate progress, review results and establish treatment regimens in a single visit (Larsson, Greig-Pylypczuk, and Huisman 2015). This may help diabetes patients getting well controlled faster and reduce the risk of patients not showing up for the test results and hence not reacting in time to poor measures. It may therefore be hypothesized that POCT improve diabetes control, patient outcomes and enhanced clinical efficiency. The equipment for POCT of HbA1c to diagnose and control type 2 diabetes exists and the equipment and materials such as test cartridges ad dilution pouches is becoming cheaper and more valid. To the best of our knowledge there is a lack of studies of the importance of POCT access to diagnostics for diabetes patients. In this study, the hypothesis is that POCT increase the quality of ambulatory care and that there is a negative link between use of POCT for HBA1c and hospital admissions for ACSCs. Therefore, the aim of this study is to describe and analyse the association between hospitalization for ACSC and use of POCT of hbA1c among diabetes patients in general practice.

Context

According to guidelines for diabetes management the HbA1c level should be measured approximately 4 times per year for a typical type 2 diabetes patient in primary care. But the exact number is depending on the individual patient's characteristics. One time a year, an annual comprehensive control to the GP should comprise a range of blood-tests and urin-analysis and EKG. The GP is rewarded for the annual control through a tariff for control of chronic diseases which the GP is allowed to use once a year per chronic condition. The remaining control visits comprise at least tests of blood percent, HbA1c and blood sugar via own equipment or external laboratories. If the GP has the required POCT equipment for HbA1c testing around 4 diabetes management visits should be sufficient per year. In this case the GP will use the following fees: blood percent (tariff 7108), Haemoglobin A1C (tariff 7403) and blood sugar (tariff 7136). If the GP does not have his own POCT equipment, a blood sample will be taken at the GP clinic and forwarded to an external laboratory a week before a second consultation with the GP. In this case the remuneration of the remaining control activities will include a tariff for a laboratory test (tariff 2101) and a standard consultation for the work of a nurse (0101) and a follow-up GP visit (tariff 0101) a week later. Therefore, lack of POCT equipment normally means more visits to the GP and delayed medical decision making.

In 2008, a special national FFS fee for point of care testing of HbA1c was agreed between the General Practitioners Organisation (PLO) and the Board for Wages and Tariffs of the Regions (RNTL) as part of a framework agreement. GPs are paid a FFS fee of € 15.49 per POCT of HbA1c (2015 prices). Each of the five Danish Regions can decide to allow their GPs to use this new tariff as an alternative to fees for standard laboratory testing and consultations. The idea is to incentivize POCT of HbA1c to secure timely and appropriate diabetes management. Nevertheless, this incentive has only been implemented by the Capital Region.

Methods:

In this study we use descriptive statistics and logit models to estimate the association between POCT of HbA1c and preventable hospital admissions including outpatient treatment. The logit model is an established method to estimate explanations for hospitalizations in the literature, see e.g., (Bottle, Aylin, and Majeed 2006) . The logit model assumes that there is an underlying unobservable risk (response variable) y defined by the relationship:

$$y_i^* = \beta' x_i + u_i \quad (1)$$

Where what we observe is a dichotomous variable y, since a patient either can be admitted (y=1) or not admitted (y=0) and characterized by a set of covariates x_i. The y variable is defined by

$$y = 1 \; if \; y > 0$$
$$y = 0 \; otherwise \quad (2)$$

In this model $\beta' x_i$ is $E(Y^*_i \mid x_i)$, and from the relations (1) an (2) we get

$$Prob(y_i = 1|x_i) = Prob(u_i > -\beta' x_i) = 1 - F(\beta' x_i) \quad (3)$$

Where F is the cumulative distribution function (or connection function) for u_i which takes values between 0 and 1. Thus $0 < F(z) < 1$ for all $z \in R$. This means that the observable values of y are realizations of a binominal process with probabilities given by (3) and changing from trial to trial depending on x_i. Thus, the likelihood function is

$$L(y_i|x_i;\beta) = \prod_{y_i=0} F(-\beta' x_i) \prod_{y_i=1} [1 - F(-\beta' x)] \quad (4)$$

Where the functional form of F() in the coproduct in (4) will depend on the assumptions made about u in (1). In this case, where the cumulated distribution of u_i is the logistic the result is the closed form expressions:

$$1 - F(-\beta' x_i) = \frac{1 + \exp(\beta' x_i)}{1 + \exp(\beta' x_i)} - \frac{1}{1 + \exp(\beta' x_i)} = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \quad (5)$$

In this study we have used alternative logit model approaches to estimate clustered dichotomous data. We have estimated standard logistic regression models with clustered data (clinics) and two level mixed-effect logistic regression model, where for a series of M independent clusters, and conditional on a set of random effects u_j:

$$Pr(y_{ij} = 1|U_j) = H(X_{ij}\beta + Z_{ij}U_j) \quad (6)$$

In (6) j =1,..., M represents clusters (GP clinics) with cluster j = 1,....,n_j observations. *The responses are the binary-valued dependent variable y_{ij} where we follow the convention of treating y_{ij} = 1 if the dependent variable $y_{ij} \neq 0$ and treating yij = 0 otherwise. The vector x_{ij} are the covariates for the fixed effects.* The vector Z_{ij} are the covariates corresponding to the random effects. They can be used to represent both random intercepts and random coefficients. The stata procedure xtmelogit and standard logit regression commands with clustering were used for Maximum likelihood estimation of the logit-estimators in stata. The marginal effects of changes in independent variables are reported as odds ratios with 95% confidence intervals. The OR estimates how much more likely it is for an

individual to be hospitalized (y=1) if the xj =1 relative to Xj =0. The BIC and AIC criteria was used to choose the best models. The variance inflation factor (VIF) and a standard regression-procedure was used to measure the extent of multicollinearity. A measure of R2 for logit models was used to estimate the explanatory power of the models (Tjur 2009).

Dependent variable

Preventable hospital admissions were defined as ambulatory care sensitive conditions (ACSCs) and recoded to a dummy variable for this event. ACSCs are conditions for which, good outpatient care can potentially prevent the need for hospital admission, or for which early intervention can prevent complications or more severe disease. This means admissions for diagnoses that could have been prevented or ameliorated with appropriate primary care. In this study we define hospitalizations in two ways. Definition 1) includes admissions (excluding outpatient activities). Definition 2) comprises all preventable admissions including outpatient activities of diabetes patients for ACSCs across the 16 different ACSC sub-categories (Schiotz et al. 2011). In both cases - patients with diabetes may be hospitalized and/or visit a hospital ambulatory (for e.g. diabetes complications) if their conditions are not adequately monitored or if they do not receive the patient education needed for appropriate self-management. But diabetes patients may also be hospitalized for other conditions such as comorbidities. Preventable hospital admissions for ACSC were not calculated by the applied algorithm for patients above 74 years. Patients above 74 were excluded.

Independent variables

The main variable of interest is whether the patient have had access to POCT. The amount and prevalence of POCT of HbA1c among diabetes patients was measured via the use of fee 7403 in the Danish Health Service Register. A dummy variable was generated to measure whether a patient received POCT of HbA1c (0= No POCT, 1 or more = POCT). Furthermore, the amount and prevalence of blood samples for central laboratory testing was measured via a FFS fee for testing via external laboratories and a generated dummy variable. Being interested in the association between POCT and ACSC hospitalization we need to control for other observable sources of variation in the risk of ACSC. This study, therefore, control for demographic, socioeconomic, health care utilization and case mix markers to adjust for patient characteristics. All continuous co-variates were divided into intervals. For instance, age was grouped into age bands. Both age and gender is expected to be positively linked to ACSC-hospitalizations. Gender and increasing age are expected to increase odds for

hospitalization. Socioeconomic markers such as income, unemployment, type of education, type of ethnicity, cohabitation status and retirement status are also expected to be linked to the likelihood of a hospitalization. Utilization of other health care services and the rural distribution classification are expected to be associated with hospitalizations. Casemix was measured via the Charlson index and measures of cost of care in primary care and secondary care. All cost covariates were grouped into trisections and the Charlson index was grouped into three groups according to the following index interval 1 [0], 2 [1-7] and 3 [8-14]. Next, the Charlson index and inpatient DRG-costs were used to adjust for morbidity burden and need of patients in secondary care. Rather than unavailable diagnoses coding from primary care, this study uses expenditures for remuneration of GP fees, drug costs, other primary care costs and outpatient DAGS-costs were used to adjust for casemix outside the hospital sector. To account for the municipality type and municipality characteristics a standard rural district classification was used (Kristensen 2006). The latter divides municipalities into four categories: Rural, peripheral, intermediate and urban municipalities.

Data:

We use register data from the year 2011 and an algorithm based on The Danish Drug Register, the Danish Health Service Register and the National Patient Register to define a population of 172,906 diabetes patients of which 11,373 were treated in GP clinics which used POCT of hbA1c equipment. The included patients were required to be above 18 and blow 75 year and comply with a standard set of criteria based on prescription of anti-diabetic drugs, blood sugar level or HbA1c test and ICD10 codes from the hospital sector. 44,981 of these patients were linked to GP clinics in the Capital Region which had more than 400 patients. Non-standard clinics with less than 400 patients were excluded. 11,373 patients of the 44,981 patients were linked to 201 GP clinics with POCT equipment and lived in the Central Region. POCT clinics were defined as clinics which have used the special fee for POCT of HbA1c at least 5 times in 2011. Thus this study analyses 11,373 diabetes patients who living in the Central Region and were listed in at GP clinics which used POCT equipment.

Results:

Table 1 shows patient characteristics for the population of 11,373 diabetes patients in the Capital Region of Denmark in 2011 who were linked to GP clinics with equipment forPOCT of HbA1c.

Table 1 Descriptive patient characteristics in GP clinics with POCT in the Capital Region 2011

	Mean	SD	CV	P5	P95
N	11,373	-	-	-	-
Demographic markers:					
Age	59.3	11.2	0.19	33	73
Distribution on age groups (%)					
20-39	7.0				
40-49	11.4				
50-59	23.8				
60-69	40.1				
70-	17.8				
Gender (male=1) (%)	0.55	0.50	0.90	0	1
Socioeconomic markers:					
Income (€)	31,485	30,380	0.96	14,638	58,901
Unemployed (%)	6.2	0.24	3.90	0	1
Short education (%)	0.34	0.47	1.39	0	1
Other ethnicity	0.20	0.40	2.02	0	1
Single (%)	0.30	0.46	1.51	0	1
Retired (%)	0.54	0.50	0.92	0	1
Health care utilisation:					
ACSC diabetes (#) - def. 1 inpatient	0.012	0.15	12.37	0	0
Total ACSCs (#) – def. 1 inpatient	0.045	0.31	6.73	0	0
ACSC diabetes (#) – def. 2 inpatient and outpatient	0.79	2.13	2.70	0	5
Total ACSCs (#) – def. 2 inpatient and outpatient	0.95	2.42	2.56	0	6
Hospital visits (#)	0.44	1.19	2.72	0	2
ED visits (#)	0.17	0.51	3.06	0	1
Number of GP visits (N_consultations) (#)	6.81	5.97	0.88	1	18
Number of POCT of HbA1c (#)	0.98	1.52	1.55	0	4
Case mix markers:					
Charlson index	0.86	1.31	1.52	0	3
Cost all GPs (€)	261.5	198.1	0,75	27.5	617.8
Other primary care costs (€)	378.8	485.0	1.28	0	1,137.3
Drug costs (€)	721.8	822.9	1.14	16.1	2,257.3
Inpatient hospital DRG-costs (€)	2,059.2	7,513.1	3.65	0	11,173.4
Outpatient hospital DAGS-costs (€)	1,781.8	5,184.1	2.91	0	5,973.6
Total health care cost including drugs (€)	5,254.1	10,482.1	2.00	382.2	18,726.1
Municipality type - rural district classification (%)					
Peripheral municipalities	1.21				
Rural municipalities	0				
Intermediate municipalities	10.80				
Urban municipalities	87.99				
Walk-in laboratory in municipality	40.0				
Testing of HbA1c					
Point of care testing D7403 – one or more (%)	41.5	49.3	1.19	0	1
Blood sample for central lab. – one or more (%)	32.4	46.8	1.44	0	1

Table 1 reveals that the mean diabetes patient in the population was 59 years with 33 years (p5) and 73 (p95) in 2011. 55% of the patients were males and the average income was € 31,485. 34% had only a short education and 20% had another ethnicity than Danish. 6.2% of the patients were unemployed and 54% were retired. Diabetes patients in this population experienced an average of 0.045 preventable hospital admissions for ACSCs (def. 1) and an average of 0.012 diabetes related

admissions for ACSCs (def. 1). The broader definition of ACSC (def. 2) including outpatient hospital activity shows that diabetes patients received an average of 0.95 hospital treatments including outpatient treatment in total and 0.79 related to their diabetes condition. In 2011, the average number of GP visits were 6.8 with (p5=1) and (p95=18) and the mean number of POCT of HbA1c was 0.98 times per patient with (P5=0 and p95=4). The latter indicates a relatively small variation (CV=1.55) in POCT and number of GP visits compared to other types of health care utilisation. The applied set of case mix markers revealed large variation in hospital morbidity. The charlson index was on average 0.86 with CV=1.52 and 0 (P5) and 3 (P95). In terms of DRG-costs (fee-weighted DRG-services) the variation in hospital case mix were even larger (CV=3.65). The total cost of care for remuneration of GP-services was the minimum of the included case mix markers based on costs and also fluctuated relatively less than all other case mix-measures based on costs. Besides expected high inpatient and outpatient hospital costs, drug costs were the second largest cost markers with relatively high variation compared to primary care costs. The total health care cost was on average € 5,254 with a median of € 2,381.7. Table 1 also reveals the geographical and age distribution of diabetes patients in the Capital Region.

Fig. 1 Variation in hospitalization admissions for ACSC (def.# 2) across municipalities

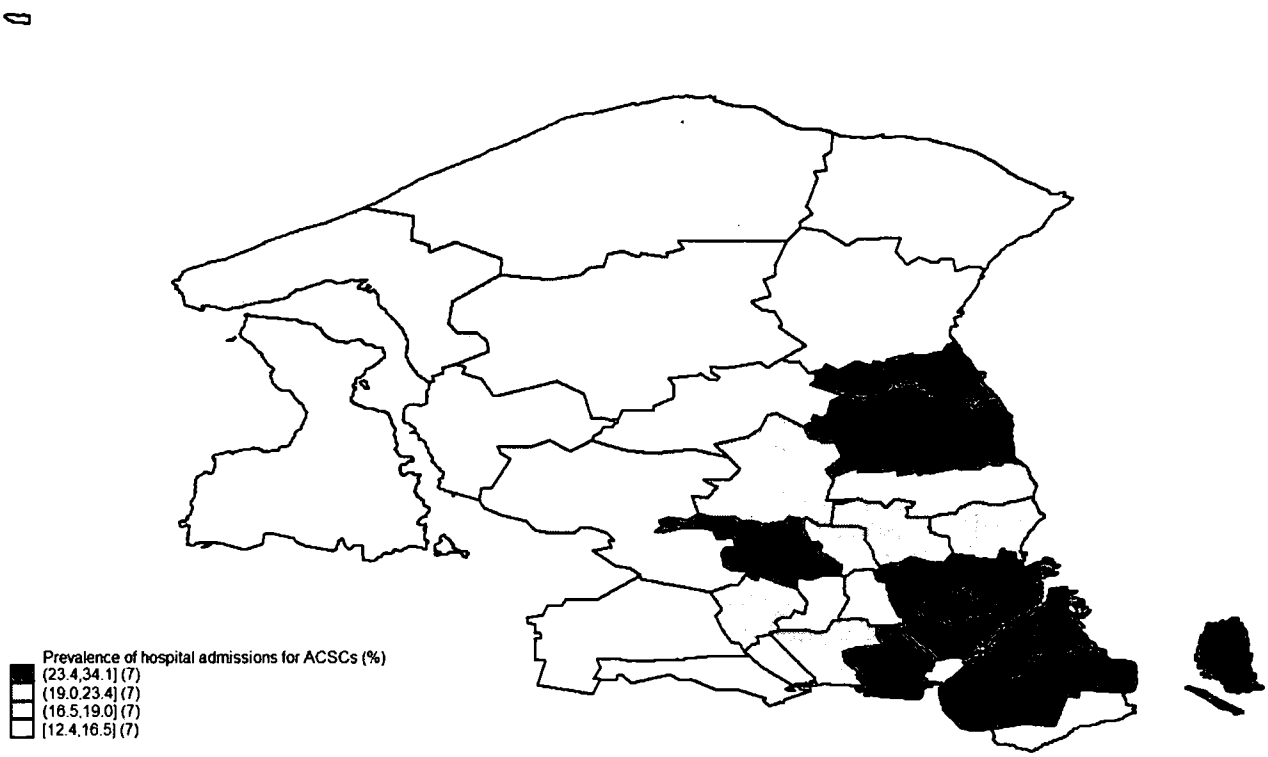


Figure 1 shows the prevalence of hospitalizations including outpatient care for ACSCs among diabetes patients in POCT clinics in the municipalities according to definition 2 in the Capital

Region in 2011. Each class corresponds to quantiles of the distribution of hospitalizations for ACSCs across municipalities, so that each class includes 7 of the 28 included municipalities. The results indicate that a larger percentage of the patients in the south-east part of the Region received preventable hospital activities than diabetes patients in the north-west part of the Region.

Table 2 Odds ratios for hospital admissions for ACSC (definition#2) and POCT of HbA1c

	Odds Ratio	P	95% conf. interval	Reference Group
Demographic makers				
Age 20-39	2.88	0.000	[2.24;3.68]	Age 50-59
Age 40-49	1.39	0.002	[1.13;1.70]	Age 50-59
Age 60-69	1.03	0.768	[0.87;1.21]	Age 50-59
Age 70-	1.03	0.802	[0.83;1.28]	Age 50-59
Gender (male=1)	1.01	0.822	[0.90;1.14]	Female=0
Socioeconomic markers				
Income (€) 2nd trisection	0.94	0.416	[0.81;1.09]	1st trisection
Income (€) 3rd trisection	0.92	0.339	[0.77;1.09]	1st trisection
Unemployed	1.01	0.923	[0.78;1.32]	Other status
Short education	1.06	0.346	[0.93;1.21]	Other status
Other ethnicity	1,37	0.000	[1.16;1.62]	Other status
Single	0.99	0.874	[0.88;1.13]	Other status
Retired	0.75	0.001	[0.63;0.89]	Other status
Case mix markers				
Charlson morbidity index[1;7]	27.42	0.000	[21.5;34.9]	Index=0
Charlson morbidity index[8;14]	18.00	0.000	[7.57;42.81]	Index=0
GP costs (€) 2nd trisection	0.88	0.090	[0.75;1.02]	1st trisection
GP costs (€) 3rd trisection	0.66	0.000	[0.56;0.79]	1st trisection
Other primary costs (€) 2nd trisection	0.98	0.837	[0.85;1.14]	1st trisection
Other primary costs (€) 3rd trisection	0.96	0,575	[0.82;1.11]	1st trisection
Total cost of drugs (€) 2nd trisection	14.28	0.000	[7.44;27.42]	1st trisection
Total cost of drugs (€) 3rd trisection	18.28	0.000	[9.49;35.19]	1st trisection
Outpatient costs (€) 2nd trisection	1.66	0.000	[1.39;1.99]	1st trisection
Outpatient costs (€) 3rd trisection	2.44	0.000	[2.04;2.92]	1st trisection
Municipality type markers				
Intermediate municipality	0.64	0.286	[0.28;1.45]	Peripheral
City municipality	0.54	0.129	[0.24;1.20]	Peripheral
Walk-in laboratory in municipality	1.60	0.000	[1.38;1.86]	No walk-in lab
Point of care testing(POCT) D7403	0.76	0.000	[0.65;0.88]	No POCT
Constant	0.002	0.000	[0.000;0.006]	
Number of patients (N)	11,373			
Number of groups	201			
R2 (Tjurs' coef. of discrimination)	0.41			
Wald Chi2 (26) =1435.76, p =	0.0000			
AIC	7048.08			
BIC	7253.57			

Variance inflation factor (mean)	2.12

Table 2 shows the adjusted odds ratios for hospitalization admissions for ACSCs based on logistic regression. Overall, the estimated proxy for explanatory power R2 (Tjurs') indicates that the included patient characteristics were able to explain 41% of the variation in the prevalence of preventable hospital activity (hospitalizations including outpatient activity) among the 11,373 patients who were linked to 201 POCT clinics in the Central Region. The individual odds ratios indicate how each of the individual markers are linked to the likelihood of a preventable hospitalization. The estimated link between diabetes patient characteristics and likelihood of preventable hospital treatment are significant for the majority of covariates. **The result indicates that there is a negative and significant link between POCT (versus NO POCT) and preventable hospital admissions (including outpatient hospital care) for diabetes patients in 2011.** One or more POCT measures of point of care measurement results in 24% lower risk of preventable hospital treatments of diabetes patients. The adjusted odds ratio for hospital treatment is significantly higher for younger patients than the older reference group (age50-59). Males with diabetes do not have higher risk than women. The Income level and short educations was not significant (p5). Patients with other ethnicity have significantly higher odds for hospital treatment than the reference group. In contrast, patients who were retired had a 25% decrease in the odds for use of POCT vs no retired.

The main part of case mix measures, which reflect elements of the morbidity burden in primary and/or secondary care, reveal that the odds for hospital treatment of ACSCs are significant. The odds for hospital treatment are significant and many times higher among patients who have a charlson index of [1;7] and [8;14] versus patients with an index of 0. This is also the case for drug cost (2nd and 3rd trisection) and outpatient costs (2nd and 3rd trisection) versus each of their reference groups. In contrast, patient who experienced a higher morbidity burden in terms of cost (3rd trisection) in general practice have significantly lower adjusted odds for hospital treatment for ACSCs versus patients in the 1str trisection of costs. The rural district classification was not linked to the likelihood of a preventable hospital treatment of ACSCS in 2011. However, the availability walk-in facilities seemed to imply increased odds for hospital treatment of ACSCs in the Capital Region in 2011.

Table 3 reveals the adjusted odds ratios for the narrow definition of total hospital admissions for ACSC (def. 1). This model shows less explanatory power. The Odds ratios indicate that younger (age20-39) and older (age70-) versus middle aged (age50-59) as well as males versus females have higher odds for hospital admissions for ACSCs. Increased morbidity burden in terms of case mix measures such as the Charlson index, GP costs and outpatient cost in hospitals also mean higher odds for hospitalization compared to the reference groups. However, POCT of HbA1c did not have

significant impact on hospitalizations for ACSCS for diabetes patients despite the fact that the magnitude of the odds indicates reduced odds.

Table 3 Odds ratios for hospital admissions for ACSC (definition#1) and POCT of HbA1c

	Odds Ratio	P	95% conf. interval	Reference Group
Demographic makers				
Age 20-39	1.78	0.011	[1.14;2.78]	Age 50-59
Age 40-49	1.28	0.273	[0.82;2.01]	Age 50-59
Age 60-69	1.31	0.128	[.93;1.85]	Age 50-59
Age 70-	1.67	0.011	[1.12;2.48]	Age 50-59
Gender (male=1)	1.41	0.003	[1.13;1.78]	Female=0
Socioeconomic markers				
Income (€) 2nd trisection	1.06	0.640	[0.82;1.38]	1st trisection
Income (€) 3rd trisection	0.75	0.086	[0.54;1.04]	1st trisection
Unemployed	1.89	0.006	[1.20;2.98]	Other status
Short education	1.28	0.032	[1.02;1.59]	Other status
Other ethnicity	1.13	0.388	[0.86;1.48]	Other status
Single	1.01	0.953	[0.78;1.30]	Other status
Retired	1.24	0.247	[0.86;1.76]	Other status
Case mix markers				
Charlson morbidity index[1;7]	10.24	0.000	[5.64;18.60]	Index=0
Charlson morbidity index[8;14]	9.50	0.004	[2.05;44.02]	Index=0
GP costs (€) 2nd trisection	1.35	0.047	[1.00;1.82]	1st trisection
GP costs (€) 3rd trisection	2.06	0.000	[1.48;2.86]	1st trisection
Other primary costs (€) 2nd trisection	1.09	0.561	[0.81;1.47]	1st trisection
Other primary costs (€) 3rd trisection	1.12	0.441	[0.84;1.48]	1st trisection
Total cost of drugs (€) 2nd trisection	1.79	0.216	[0.71;4.47]	1st trisection
Total cost of drugs (€) 3rd trisection	3.29	0.013	[1.28;8.45]	1st trisection
Outpatient costs (€) 2nd trisection	1.54	0.020	[1.07;2.24]	1st trisection
Outpatient costs (€) 3rd trisection	1.73	0.002	[1.23;2.44]	1st trisection
Municipality type markers				
City municipality	1.44	0.081	[0.96;2.17]	Non rural
Walk-in laboratory in municipality	1.26	0.057	[.99;1.59]	No walk-in lab
Point of care testing(POCT) D7403	0.91	0.434	[0.69;1.17]	No POCT
Constant	0.00	0.000	[0.00;0.00]	
Number of patients (N)	11,373			
Number of groups	201			
R2 (Tjurs' coef. of discrimination)	0.062			
Pseudo R2	0.1836			
Wald Chi2 (25) =449.73, p =	0.0000			
AIC	2704.6			
BIC	2895.4			
Variance inflation factor (mean)	2.33			

Discussion:

This study did not reveal a link between POCT and outcome in terms of preventable hospital admissions for the narrow definition of ACSC without outpatient hospital activities (def.#1). This result is in line with the literature where there is an absence of evidence of the effectiveness of POCT for HbA1c in the management of diabetes (Al-Ansary et al. 2011). Our literature search did not reveal previous studies about this link. The missing association may be related to the fact that patients that access specialist care perceive that GPs have a limited role in their disease management. Although diabetes patients may have good access to care, they have also identified several factors outside the scope of general practice management (Manski-Nankervis et al. 2015). Thus, it can be argued that diabetes patients' admission to hospital may not have been avoidable. However, this study indicates that there was a significant link between POCT of HbA1c in general practice and the broader definition of ACSC including outpatient hospital activity (def. #2). Patients that experienced good access to care in terms of POCT seem to have lower odds for outpatient services at hospitals.

The applied approach in this paper has limitation. The GPs may choose to use POCT on a group of patients who has special and unobserved characteristics. For instance, GPs may choose to use POCT for patients who are more compliant (unobserved) or has a lower diabetes age (unobserved) than other diabetes patients. Apart of the patients who do not get POCT may have been tested via a standard blood sample for laboratory testing or through special walk-in laboratory facilities in the Central Region. Still, another part of the patients who do not get POCT may be monitored in hospital ambulatories or treated via "shared care" agreements between GPs and hospitals. The latter circumstances may explain at least a part of the association between POCT and the broader definition of hospital admissions for ACSCs including outpatient activity. Thus, there is a risk that this part of the analyses has revealed an obvious relationship. To address these issues other approaches and methods (e.g. propensity score matching) may be relevant.

We assume that POCT equipment contributes to improved quality of care compared to external laboratory testing. However, this assumption can be questioned because of potential lack of organization of quality assurance of equipment and procedures. The GPs must make sure that their equipment is controlled on a current basis to secure that their measurements can match the laboratory service that physicians receive at hospitals. This requires an effort and time, which the GP may decide to use for other purpose. Therefore and because clinics often have to use external laboratory testing for other laboratory values than HbA1c they may prefer to use external laboratory tests rather than testing of Hba1c on their own equipment. The latter also allow GPs to minimize

their responsibility for inappropriate quality assurance of own equipment and related procedures. Consequently, use of external laboratory testing does not automatically imply lower quality of care (or vice versa).

Conclusion:

Our preliminary results indicate that more POCT of HbA1c may result in fewer preventable hospital treatments for ACSCs among diabetes patients in general practice.

Literature

Al-Ansary, L., A. Farmer, J. Hirst, N. Roberts, P. Glasziou, R. Perera, and C. P. Price. 2011. "Point-of-care testing for Hb A1c in the management of diabetes: a systematic review and metaanalysis." *Clin Chem* 57(4): 568-76.

Ansari, Z., N. Carson, A. Serraglio, T. Barbetti, and F. Cicuttini. 2002. "The Victorian Ambulatory Care Sensitive Conditions study: reducing demand on hospital services in Victoria." *Aust Health Rev* 25(2): 71-7.

Arrieta, A. and A. Garcia-Prado. 2015. "Cost sharing and hospitalizations for ambulatory care sensitive conditions." *Soc Sci Med* 124: 115-20.

Bottle, A., P. Aylin, and A. Majeed. 2006. "Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis." *J R Soc Med* 99(8): 406-14.

Connell, F. A. 1985. "Epidemiologic approaches to the identification of problems in diabetes care." *Diabetes Care* 8 Suppl 1: 82-6.

Gao, J., E. Moran, Y. F. Li, and P. L. Almenoff. 2014. "Predicting potentially avoidable hospitalizations." *Med Care* 52(2): 164-71.

Kristensen, I. T., Kjeldsen, C. og Dalgaard, T. 2006. "Landdistrikskommuner - indiaktorer for landdistrikt." Danmarks Jordbrugsforskning. Afdeling for Jordbrugsproduktion og Miljø, Tjele.

Larsson, A., R. Greig-Pylypczuk, and A. Huisman. 2015. "The state of point-of-care testing: a European perspective." *Ups J Med Sci* 120(1): 1-10.

Manski-Nankervis, J. A., J. Furler, R. Audehm, I. Blackberry, and D. Young. 2015. "Potentially preventable hospitalisations: are they a useful marker of access to and experience of care in general practice among people with type 2 diabetes?" *Aust J Prim Health* 21(2): 214-20.

Moy, E., E. Chang, M. Barrett, C. Centers for Disease, and Prevention. 2013. "Potentially preventable hospitalizations - United States, 2001-2009." *MMWR Surveill Summ* 62 Suppl 3: 139-43.

Schiotz, M., M. Price, A. Frolich, J. Sogaard, J. K. Kristensen, A. Krasnik, M. N. Ross, F. Diderichsen, and J. Hsu. 2011. "Something is amiss in Denmark: A comparison of preventable hospitalisations and readmissions for chronic medical conditions in the Danish Healthcare system and Kaiser Permanente." *Bmc Health Services Research* 11.

Tjur, T. 2009. "Coefficients of Determination in Logistic Regression Models-A New Proposal: The Coefficient of Discrimination." *American Statistician* 63(4): 366-72.

Diabetes related deaths in The Danish malnutrition period 1999-2007

Maja Sparre-Sørensen *University Hospital Herlev, geriatric section 1*
Gustav Kristensen *Stockholm School of Economics, Latvia*

Abstract.

Background: Over the last few years scientist have become aware of what is known as the obesity paradox. Thou obesity in its own right increases the risk of developing type 2 diabetes, it seems that patients suffering from type 2 diabetes have a better survival rate if they are overweight or moderately obese and weight stabile, compared to leaner type 2 diabetes patients and/ or type 2 diabetes patients who have lost weight.
Studies have also shown that patients suffering from type 1 diabetes also have an increased mortality rate if they are underweight or lose weight. During the Danish malnutrition period from January 1999 to January 2007, there was a statistically significant increase in the number of deaths related to malnutrition among the elderly in Denmark. Many more may have been suffering from malnutrition, but not to such a degree that it led to their deaths.
Objective: The aim of this study is to examine whether or not the effect of the malnutrition period can be seen in the number of diabetes-related deaths.
Method: Regression analyses based on the Expansion Method.
Results: We found a sudden statistically significant rise in the number of deaths from diabetes associated to the period when the general nutritional state among the elderly in Denmark worsened (from 1999 to 2007).
Cohorts plays an important role in the development of the death rate from diabetes.
An association between the Danish malnutrition period 1999 – 2007 and the death rate from all-cause diabetes in Denmark could be confirmed.

Conclusion: The study concludes that the malnutrition period resulted in an excess death rate from all-type diabetes. In total, approximately 1216 extra lives were lost during this period to diabetes.

Keywords: diabetes, obesity paradox, cohorts, malnutrition, death rates, The Danish malnutrition period.

[1] Correspondence address: Maja Sparre-Sørensen MD, University Hospital Herlev, geriatric section, Denmark. E-mail: majasparre@gmail.com

Introduction

Over the last few years scientist have become aware of what is known as the obesity paradox. Thou overweight in its own right increases the risk of developing of type 2 diabetes, it seems that patients suffering from type 2 diabetes have a better survival rate and fewer non-fatal cardiovascular events, if they are overweight or moderately obese and weight stabile. Although weight loss has been recommended for patients suffering from type 2 diabetes it now seems that health professional need to reevaluate this praxis [1-2]. Patients suffering from type 1 diabetes also have to be aware of weight loss and being underweight as this increases mortality and the risk of developing ketoacidosis [3-4]. Ketoacidosis is a serious complication to type 1 diabetes, and is lethal in 1-5% of all cases. The risk of dying from ketoacidosis increases with age and comorbidity [5].

During the Danish malnutrition period from 1999 to 2007, mortality rate from malnutrition increased and 8 years later decreased rapidly practically form one day to the next. This caused approximately 341 extra deaths as a direct result of malnutrition and almost 3900 extra deaths from stroke, schizophrenia and Alzheimer's disease [6-10].

Inspired by the Dutch famine studies [11] the purpose of this study is to model the death rate from diabetes in the period 1994-2013 and to document a derived effect from malnutrition on the death rate from all-cause diabetes.

Data and Method

Definition: Death rate is the number of death from a certain cause per 100 000 persons in a considered group. (E.g. the number of 80-84 year old women death from stroke per 100 000 women in the age group 80-84.)

The Danish data on the *death rate* from malnutrition are taken from: The State Serum Institute: Malnutrition B-040 and Diabetes (Sukkersyge) B-039 [12].

2012. The explaining variables are:

Diab	death rate from all-cause diabetes.
Dmal	death rate from malnutrition
T	the period (or year), 1977 = 1
Age	age at death
CohYear	diagonal dummy system indicating year of birth
D2000m	dummy for the year 2000, applied for men.
D2001w	dummy for the year 2001, applied for women.
B	beta coefficient vector
Gender (index)	*m* – men *w* – women

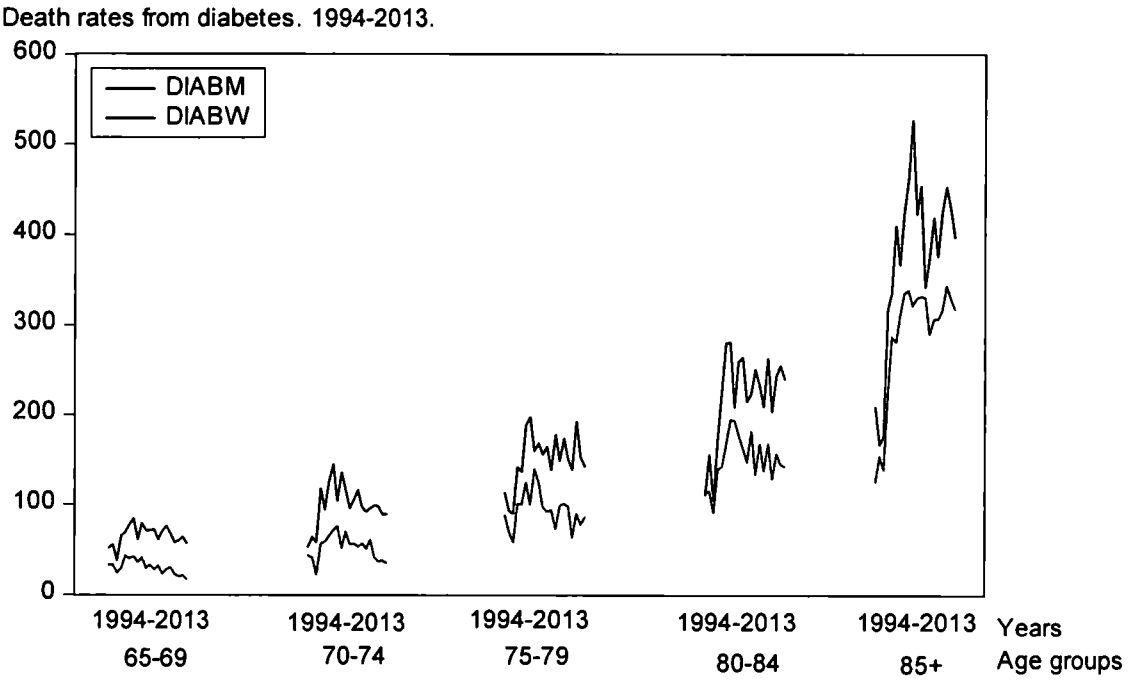


Figure 1. Death rates from diabetes 1994-2013.

Figure 1 shows the death rate from diabetes 1994-2013. The death rate is higher for men than for women. It is also seen that the death rate grows (almost exponentially) with Age.

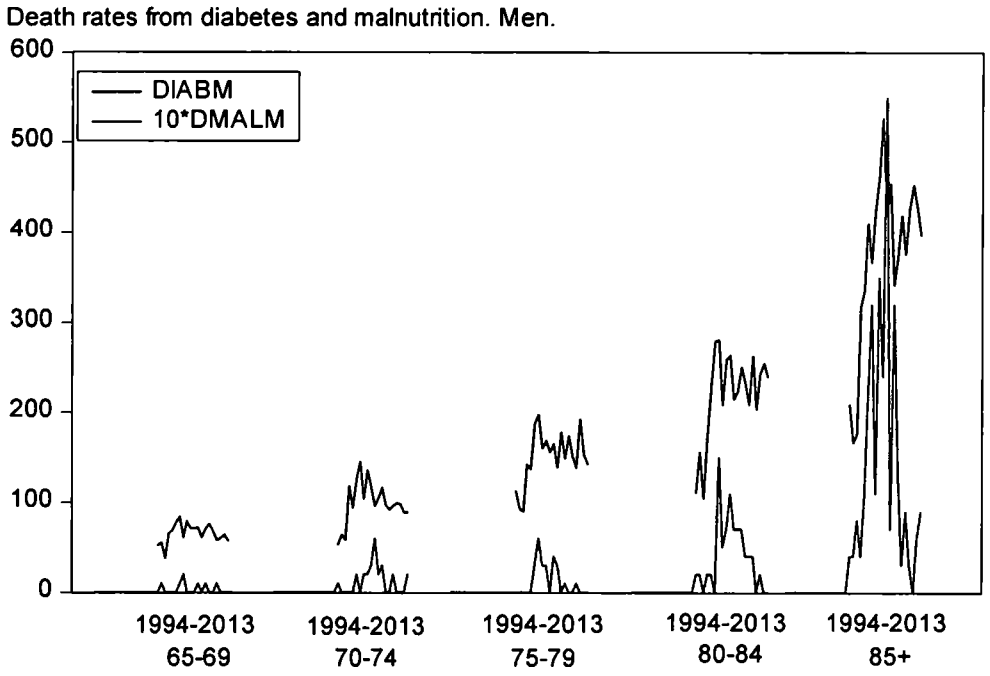


Figure 2. Death rates from diabetes and malnutrition 1994-2013.

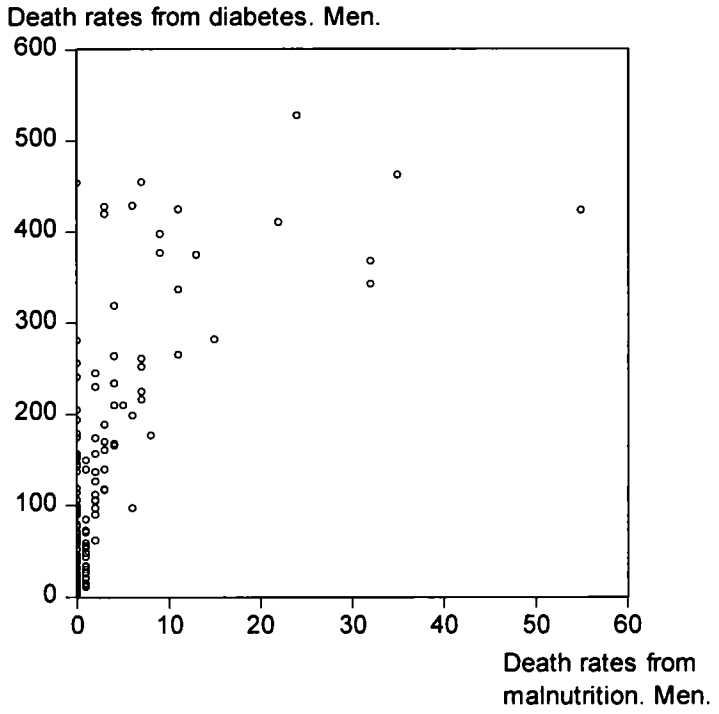


Figure 3. Death rates from diabetes and malnutrition 1994-2013.

Figures 2 and 3 gives the impression that the death rate from diabetes is associated with the death rate from malnutrition.

The choc-effect. Men and women

Men and women have a different structure in relation to the pattern of diabetes. Women appear to be more robust than men. That has an impact of the force and the time-lag by which diabetes develop for the two gender.

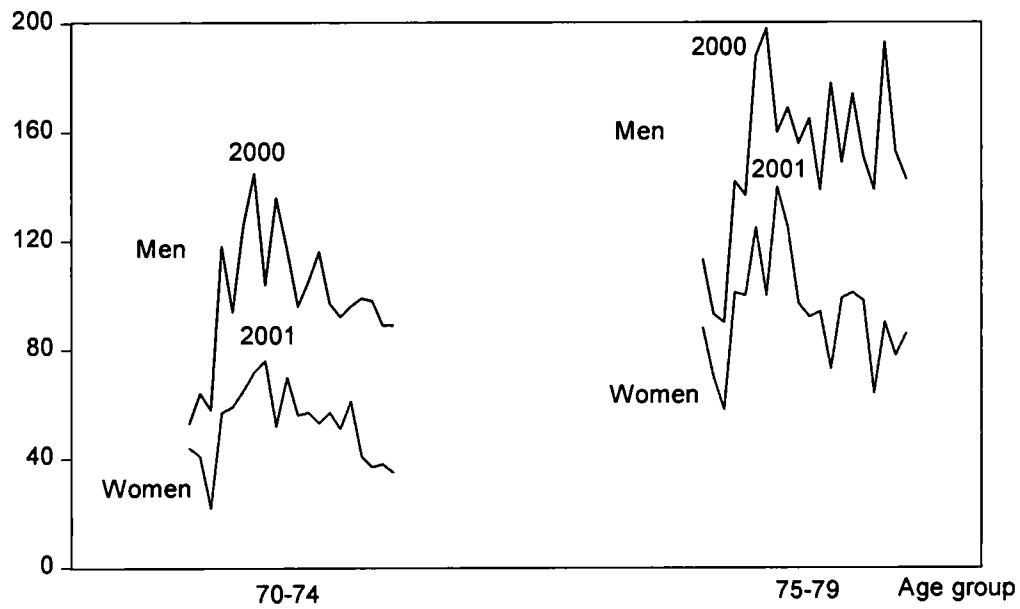


Figure 4. Death rates from diabetes top in year 2000 for men, versus 2001 for women. Women's top thus have a one year delay related to men.

Because the peak in year 2000 for men and 2001 for women we see this as a choc-effect of the malnutrition period 1999-2007.
We have now shown the different elements in death rates from diabetes and can combine the in an initial model.

Models
The steps are calculated for men and women. The models are build up in three steps:

A. Simple linear model.

$$\text{Log}(Diab) = \alpha_1 + \alpha_2 Age + \alpha_3 Age^2 + \alpha_4 Sqr(T) + \alpha_5 Sqr(Dmal) + \alpha_6 D2000 + \text{er} \quad (1)$$

T is a time trend indicating changing life style and medical progress in relation to diabetes. It is seen that it is growing over time in this model.

B. The Age-Period-Cohort model
The assumption behind the cohort dummies is that each cohort throughout life remain in the same health group, e.g., the 85-year-old in 1983 is in the same health group as the 80-year-old in 1978. The residual "er" in Equation (1) is now explained as the cohort effect [13-19].

$$\text{Log}(Diab) = \alpha_1 + \alpha_2 Age + \alpha_3 Age^2 + \alpha_4 Sqr(T) + \alpha_5 Sqr(Dmal) + \alpha_6 D2000 + \beta_{10} Coh1910 + \beta_{11} Coh1911 + \ldots\ldots + \beta_{67} Coh1967 \quad (2)$$

The cohort effect indicate that the basis for diabetes was created early in life.

Calculated by Weighted Least Square using "Age" as weight. We here for a moment disregard the remaining random element.

For each age group we have a part of the beta string as (a column vector) e.g.

$$B_{80} = [\beta_{14} \ \beta_{15} \ \beta_{16} \ldots\ldots \ \beta_{23}] \quad (3)$$

and for the entire equation:

$$B = [B_{55} \ B_{60} \ B_{65} \ B_{70} \ B_{75} \ B_{80} \ B_{85}] \quad (4)$$

The cohort effects show a big jumps upwards from 1911, and downwards from 1939 to 1949, where the Second World War and changing life style reduce the beta-coefficients. From 1949 the new generations again grows weaker in relation to diabetes.

We now have calculated a model that describes the development in the death rate from diabetes 1994-2013. The estimated set of beta coefficients can be included in equation (1) as an extra explaining variable which increase the number of degrees of freedom in relation to equation (2).

The final step is a recalculated model.

$$\text{Log}(Diab) = \alpha_1 + \alpha_2 Age + \alpha_3 Age^2 + \alpha_4 Sqr(T) + \alpha_5 Sqr(Dmal) + \alpha_6 D2000 + \alpha_7 B + \gamma_1 \epsilon(-1) \quad (3)$$

This gives a good model for the time series of death rate from diabetes.

Did malnutrition provoke increasing death rate from diabetes?
We will now test if the Danish malnutrition in the period 1999-2007 [6-10] had an impact on the death rate from diabetes.
The result of the empirical calculation became for men:

Log(Diabm) = -3.67 + .1164*Age - .000207*Age^2
(-6.69) (7.11) (-1.65)

+ .1316*Sqr(T) + .04965Sqr(Dmalm) +.2192*D2000m + .9847*B_m + .004183*ϵ(-1)
(4.55) (3.02) (3.45) (8.34) (3.27)

R^2= .9916 Adj.R^2= .9912 DW= 1.98 Obs = 169

And for women:

Log(Diabw) = -.4143 + .000657*Age^2
(-2.49) (36.92)

+ .10004*Sqr(T) + .05679(Dmalw) +.3920*D2001w + .9533*B_w
(3.27) (3.85) (5.32) (12.59)

R^2= .9898 Adj.R^2= .9895 DW= 1.86 Obs = 178

We see that the (squared) variables Dmalm and Dmalw are highly significant, which indicates that malnutrition is highly associated with the death rate from diabetes. Likewise for the dummies D2000m and D2001w.
By using those results we can calculate the effect of the malnutrition on the death rate from diabetes. The results are shown below in Table 1. for the number of deaths.

The actual values of death rates from diabetes are compared with the model calculated values in Figure 5.

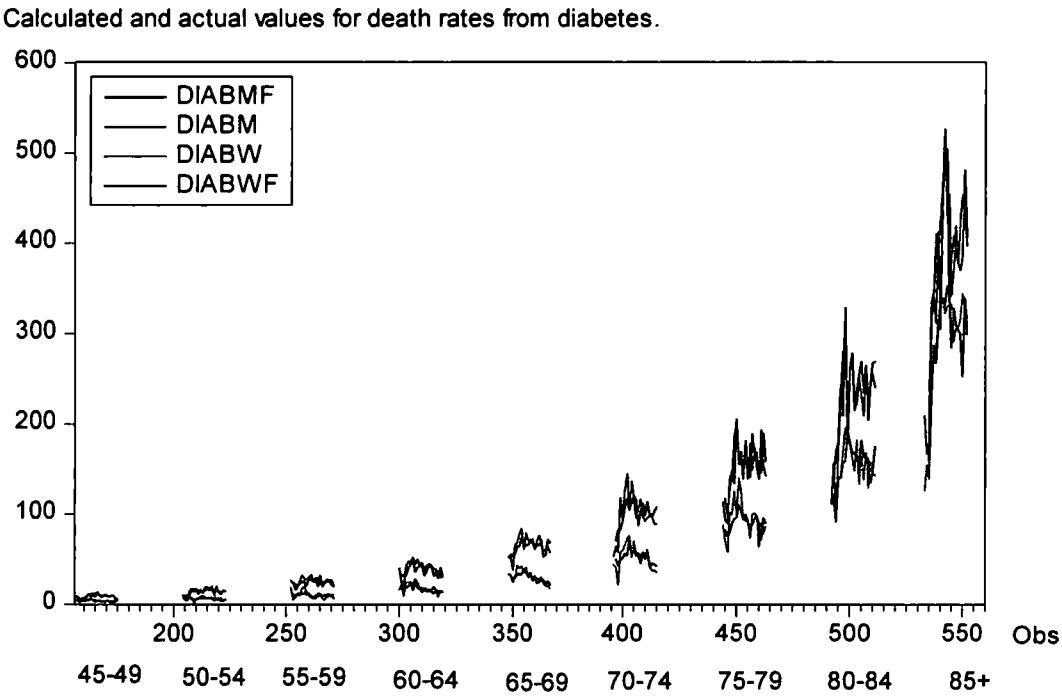


Figure 5. Calculated and actual death rates from diabetes.

The excess death rates from diabetes

By using the estimated models of equation (3) we can now calculate the excess death rates from diabetes provoked by malnutrition. The "normal" death rate from malnutrition is fixed to be the level of death rate before 1999. By inserting the "normal" death rates for malnutrition in the estimated equations for the vales Dmal we can calculate a "normal" death rate from diabetes. The difference is the "excess death rate from diabetes" associated with the excess death rate from malnutrition.

Figure 6 shows the calculated excess death rates from diabetes calculated for the age groups 70-74 and 75-79.

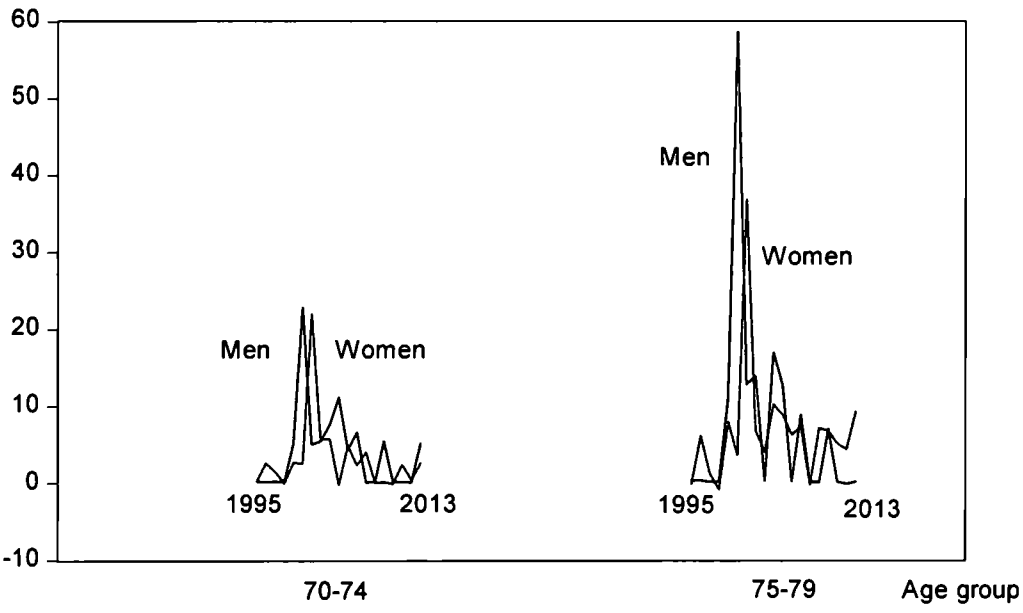


Figure 6. Calculated excess death rates from diabetes.

Number of deaths

Table 1. Number of deaths from diabetes provoked by malnutrition		
Age group	**Men**	**Women**
45-49	5.36	4.48
50-54	2.72	.54
55-59	16.82	9.17
60-64	21.06	14.82
65-69	24.29	24.32
70-74	59.08	57.05
75-79	92.36	94.73
80-84	88.67	111.09
85+	144.16	445.23
Total	**454.52**	**761.43**

Table 1 show that the death rate from diabetes during the Danish malnutrition period is highest among the old. As there are more women among the oldest of the old, more women the men die from diabetes during the Danish malnutrition period.

Discussion

Underweight and weight loss increases mortality in both type 1 and type 2 diabetes patients but in two different ways. We were in our research not able to distinguish between typel and type 2 related deaths, we therefore cannot say with certainty whether the extra deaths were caused by an increase of ketoacidosis related deaths among patients suffering from type 1 diabetes or by cardiovascular diseases among patients suffering from type 2 diabetes [1-5].

Likewise we cannot say for certain that the patients who died from diabetes in excess were malnourished.

We cannot say for sure why women seem to be more robust than men, the chock in 2000/2001 is unexplained.

A weakness in this study is that it is based on annual data. The difference in the lag structure for men and women in the modelling underline a need for applying monthly data. A number of alternative studies were made. For the purpose of this study the differences were only cosmetic.

The number of people who died directly or indirectly from malnutrition during the period 1999-2007.

Table 2. Calculated number of deaths directly or from other diseases due to malnutrition

deaths	malnutrition	diabetes	Stroke*	Alzheimer's*	Schizophrenia*
women	220	761	2106	266	138
men	121	455	1248	79	52
Total	341	1216	3354	345	190

* Based on separate studies. (6-10)

Conclusion

Malnutrition in the period 1999-2007 indirectly caused an increased death rate from all-cause diabetes related deaths. In total 1216 extra lives were lost to diabetes.

References.

1. Jeng-Fu K, Yi-Ting H, I-Chieh M MD, Shi-Dou L, Shih-Te T, Ming-Chia H. The association between body mass index and all-cause mortality in patients with type 2 diabetis melnitus, Medicine volume 94, number 34, august 2015.
2. Wolfram W, Erdmann E, Cairns R, Clark AL, Dormandy JA, Ferrannini E, Anker SD. Inverse relation of body weight and weight change with mortality and morbidity in patients with type 2 diabetes and cardiovascular co-morbidity: An analysis of the PROactive study population, International Journal of Cardiology, 162, 20-26 (2012)
3. Secrest AM, Becker DJ, Kelsey SF, LaPorte RE, Orchard TJ. Characterizing sudden death and dead-in-bed syndrome in Type 1 diabetes: Analysis from 2 childhood-onset Type 1 diabetes registries, Diabetic Medicine, 28(3) 293-300 march 2011.
4. Usher-Smith JA, Thompson MJ, Sharp SJ, Walter FM. Factors associated with the presence of diabetic ketoacidosis at diagnosis of diabetes in children and young adults: a systematic review, British Medical Journal 343:d4092, 2011
5. Westerberg DP. Diabetic Ketoacidosis: Evaluation and Treatment, American Family Physician, 87(5):337-346. 2013
6. Sparre-Sørensen M, and Kristensen GN (2015) Malnutrition related deaths. (Paper in progress).
7. Sparre-Sørensen M, and Kristensen GN (2015). Underernæringen i Danmark. 1999 – 2007. Hvordan tæller man de døde? *Symposium i anvendt statistik. Danmarks Tekniske Højskole*, 2015; 200-211.
8. Sparre-Sørensen M, Kristensen GN (2015) Alzheimer's disease in the Danish malnutrition period, 1999 – 2007. *Journal of Alzheimer's Disease*. **48** (4), 979-985.
9. Sparre-Sørensen M, Kristensen GN (2015) Schizophrenia and the Danish malnutrition period, 1999 – 2007. (Paper in progress)
10. Kristensen GN, Sparre-Sørensen M. (2015). Death from Stroke during the Danish malnutrition period 1999-2007. *Journal of Statistical and Econometric Methods*, 4, (2): 127-154.
11. Ekamper et all. (2013) Independent and additive association of prenatal famine exposure and intermediary life conditions with adult mortality between age 18-63 years, *Social science and medicine*, http://dx.doi.org/10.1016/j.socsimed.2013.10.027
12. Statens Serum Institut (2014). *Database*. www.ssi.dk
13. Kristensen GN (2014b). The Holford puzzle. The cohort effects in death rates from lung cancer. *International Journal of Health Research and Innovation.* **2** (1): 1-14.
14. Holford TR (1991). Understanding the effects of age, period, and cohort on incidence and mortality. *Annu. Rev. Publ. Health* 12: 425-457.
15. Kristensen GN (2013) Cohort Coefficients. Describing the secular development in protective and detrimental cohort effects associated with apoplexy. *Journal of Statistical and Econometric Methods.* **2**, (4): 119-127.
16. Glenn ND (1976). Cohort Analysts' Futile Quest: Statistical Attempts to Separate Age, Period and Cohort Effects. *American Sociological Review*, 41, (5): 900-904.
17. Clayton D, Schifflers E. (1987). Models for temporal variation in cancer rates. II. The age-period-cohort model. *Statistics in Medicine,* 6: 469-481.
18. Kristensen GN (2014a). Testing 'Clemmensen's hook' in the death rate from breast cancer. *Journal of Statistical and Econometric Methods.* **3,** (2): (2014), 15-30
19. Osmond C, and Gardner MJ (1982). Age, Period and Cohort Models applied to Cancer Mortality Rates. *Statistics in Medicine*, 1: 245-259.

Appendix: "The malnutrition choc"

The malnutrition period was by Sparre-Sørensen and Kristensen [2015] localized to the years 1999-2007, and defined as a period of excess mortality under the diagnosis "malnutrition" (Category B-040, underernæring og fejlernæring).

"The malnutrition choc" is here the unexpected high mortality from diabetes which is seen *after* an overall correction for the excess death rate associated with malnutrition. The choc which hit in the same year for all age groups is shown in Figures 1A and 2A

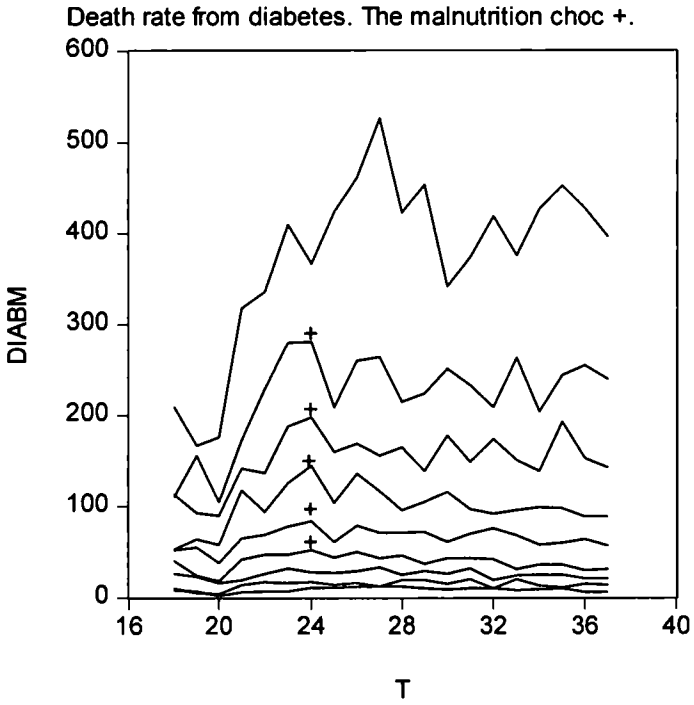


Figure 1A. The diabetes choc. Men.
T = 24 is year 2000. The "choc" might come earlier at increasing age.

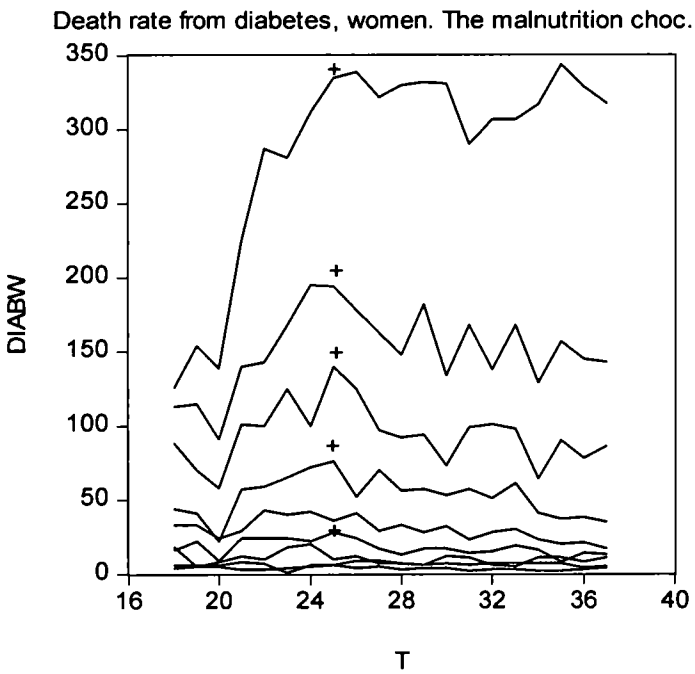


Figure 2A. The diabetes choc. Women
T = 25 is year 2001. The "choc" might come later at increasing age.

Analysis and visualisation of spatial and spatio-temporal data

Søren Lophaven
Center of Epidemiology and Screening, Department of Public Health
University of Copenhagen

1 Introduction

Spatial data are data that are linked to geographical locations, and hence can be presented on a map. Most statistical methods for analysing spatial data operate under the premise that data collected at locations close together tend to be more alike than data collected at locations further apart. Much of the methodology developed for analysing spatial data mimics that of analysing time series data (data correlated over time), where the data have a natural temporal ordering. However, for spatial data no such ordering is generally present, and this prevents a straightforward extension of time series methods to spatial data. Cressie (1993) separated spatial data into three classes: 1) spatially continuous (geostatistical) data, 2) area (lattice) data, and 3) spatial point process data. We present the different types of spatial data and show how they can be presented on a modern map.

Spatio-temporal data are data that are linked to locations in both space and time, and hence can be presented as time series for a specific location in space, on a map for a specific point in time, or as a temporal sequence of maps using an animation.

2 Examples

The first dataset contains measurements of the concentration of lead and Polycyclic Aromatic Hydrocarbons (PAH) from soil samples taken at a total of 171 locations at Østerbro. Originally, data from 138 locations were available. Subsequently, 33 additional samples were obtained from locations close to other locations already sampled. We show how the data can be

analysed and presented on a map.

The second example deals with data from Kattegat, the sea area between Denmark and Sweden. The dataset consists of surface concentrations of nutrients and phytoplankton biomass as well as salinity and temperature representing the upper 10 meters of the water column. Measurements were taken at 70 monitoring stations during the period 1993-1997, and the temporal resolution varies a lot from station to station as seen in Figure 1.

We show how the concentration of dissolved inorganic nitrogen can be modelled in space and time and subsequently presented using animations.

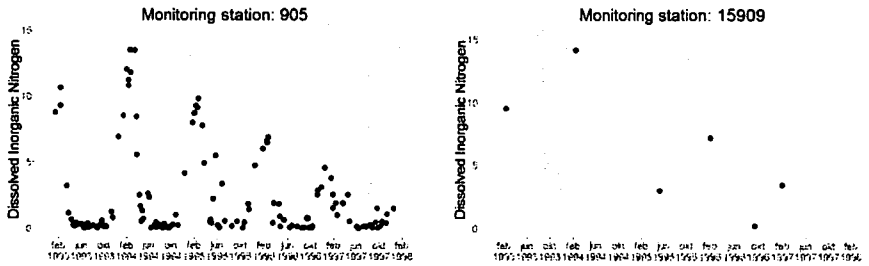


Figure 1: *The concentration of Dissolved Inorganic Nitrogen in the period 1993-1997 at two different monitoring stations in Kattegat*

References

Cressie, N. (1993). *Statistics for spatial data.* Wiley

Værdikort baseret på NFA's spørgeskemaundersøgelse "Arbejdsmiljø og Helbred i Danmark i 2014"

Hans Bay
hba@nrcwe.dk

Det Nationale Forskningscenter for Arbejdsmiljø (NFA) gennemførte i 2014 undersøgelsen "Arbejdsmiljø og Helbred " (AH 2014), som er baseret på en stikprøve af befolkningen på ca. 50.000 beskæftigede lønmodtagere i Danmark mellem 18 og 64 år. Disse personer fik i marts 2014 tilsendt en invitation til at deltage i spørgeskemaundersøgelsen 'Arbejdsmiljø og helbred i Danmark'. Spørgeskemaet omfattede 51 hovedspørgsmål (omfattende ca. 220 variable) om psykisk og fysisk arbejdsmiljø og helbred. Undersøgelsen er en gentagelse af en tilsvarende undersøgelse foretaget i 2012.

Tabel 1. Oversigt over AH undersøgelserne i 2012, 2014 og 2016

år	Brutto brutto stikprøve	datasæt
2012	35.000	16.437
2014	50.000	27.812
2016	50.000	xx.xxx

Det forventes at AH 2016 bliver af samme størrelse som den i 2014. I begge datasæt er der efterfølgende udarbejdet vægte til. I de kommende analyser er disse vægte ikke brugt.

Konstruktion af kort

Der tages udgangspunkt i datasættet fra 2014 bestående af 27.812 besvarelser. Ønsket er, at finde et begrænset antal variable, der med rimelighed kan beskrives med to underliggende faktorer/dimensioner. Efter et par indledende faktoranalyser udvælges følgende 12 variable.

Tabel 2. 12 udvalgte spørgsmål fra AH 2014

Spørgsmålene er som hovedregel besvaret på en skala fra 1 til 5:

altid	ofte	sommetider	sjældent	aldrig	har ingen leder
1	2	3	4	5	-

Nr.	spørgsmålsformulering	antal	gns.
1	Hvor ofte ... forklarer din nærmeste leder dig virksomhedens mål, så du forstår, hvad de betyder for dine opgaver?	25.977	2,64
2	Hvor ofte ... tager din nærmeste leder sig tid til at engagere sig i din faglige udvikling?	25.966	2,89
3	Hvor ofte ... involverer din nærmeste leder dig i tilrettelæggelsen af dit arbejde?	25.953	2,84
4	Hvor ofte ... giver din nærmeste leder dig den nødvendige feedback (ris og ros) for dit arbejde?	25.977	2,83

5	Hvor ofte ... bliver dit arbejde anerkendt og påskønnet af ledelsen?	25.970	2,73
6	Hvor ofte ... får du den hjælp og støtte, du har brug for fra din nærmeste leder?	25.981	2,47
7	Hvor ofte ... kan man stole på de udmeldinger, der kommer fra ledelsen?	25.969	2,33
8	Hvor ofte ... får du den vejledning og instruktion, du behøver for at udføre dit arbejde?	26.668	2,27
9	Hvor ofte... er det nødvendigt at holde et højt arbejdstempo?	26.702	2,14
10	Hvor ofte... har du tidsfrister, som er svære at holde?	26.695	2,84
11	Hvor ofte... får du uventede arbejdsopgaver, der sætter dig under tidspres?	26.702	2,79
12	Hvor ofte... oplever du, at du har nok tid til dine arbejdsopgaver?	26.704	2,64

Ovenstående spørgsmål kan delvis findes i det Svenske WOLF study. (Leisure time, occupational and household physical activity, and risk factors for cardiovascular disease in working men and women). Og delvis findes i The Copenhagen Psychosocial questionnaire (COPSOQ). (http://www.arbejdsmiljoforskning.dk/)

De 12 spørgsmål omfatter helt klart ledelse og noget om opgavernes kompleksitet og omfang. Spørgsmål om eksempelvis søvn og engagement er ikke med. Ikke overraskende viser den valgte faktoranalyse, at man kan konstruere to dimensioner. To af faktorerne har en egenværdi der er større end 1, og den kumulerede varians udgør 60 %. Den første faktor (factor1) kaldes ledelsesstøtte og den anden faktor (factor2) kaldes opgavepres. De to faktorer er konstrueret således, at bliver uafhængige standardiserede normalfordelinger. En positiv værdi for factor1 betyder lidt lederstøtte. En positiv værdi for factor2 betyder stort opgavepres. De to faktorer danner et traditionelt koordinatsystem. Hvor factor1 er x-aksen og factor2 er y-aksen.

Tabel 3. Beskrivelse af værdikortets kvadranter

Factor1	Facrot2	kvadrant	beskrivelse	Antal respondenter
factor1=>0	factor2=>0	Kvad=1	Respondenter oplever beskeden lederstøtte og stort arbejdspres	6.356
factor1<0	factor2=>0	Kvad=2	Der opleves høj støtte fra ledelsen og samtidigt stort arbejdspres	6.426
factor1<0	factor2<0	Kvad=3	Der opleves høj støtte fra ledelsen og behersket arbejdspres	8.726
factor1=>0	factor2<0	Kvad=4	Respondenter oplever beskeden lederstøtte og samtidig et beskedent arbejdspres	6.304

Konstruktion af værdier

Blandt de øvrige spørgsmål konstrueres nu 6 værdivariable (6 værdiskalaer): Engagement i arbejdet, egen kontrol, søvn, energi, stress og nervøsitet. Cronbach alpha bruges til at vurdere om spørgsmålene kan slås sammen til en skala. Selve værdivariablen/skalaen er et gennemsnit af de indgående variable.

Tabel 4. **Engagement værdien**: Cronbach alpha: 0,90. Lille værdi er udtryk for stort engagement.

1	I hvilken grad ...	giver dit arbejde dig selvtillid og arbejdsglæde?
2	I hvilken grad ...	synes du, dine arbejdsopgaver er interessante og inspirerende?
3	I hvilken grad ...	er dit arbejde vigtigt for dig (ud over indkomsten)?
4	I hvilken grad ...	føler du dig veloplagt, når du er på arbejde?
5	I hvilken grad ...	bliver du opslugt af dine arbejdsopgaver?

Engagement skala kilde ukendt.

Tabel 5. **Kontrol værdien**: Cronbach alpha: 0,66. Høj værdi udtryk for kontrol med "livet".

1	Hvor ofte inden for den sidste måned har du ...	følt, at du var ude af stand til at styre vigtige ting i dit liv?
2	Hvor ofte inden for den sidste måned har du ...	følt, at vanskelighederne på arbejdet hobede sig sådan op, at du ikke kunne klare dem?
3	Hvor ofte inden for den sidste måned har du ...	følt dig sikker på dine evner til at klare vanskeligheder på arbejdet?
4	Hvor ofte inden for den sidste måned har du ...	følt, at tingene på arbejdet udviklede sig, som du ønskede det?

Ovenstående spørgsmål er inspireret fra PSS-4 'perceived stress scale'

1. In the last month, how often have you felt that you were unable to control the important things in your life?
2. In the last month, how often have you felt confident about your ability to handle your personal problems?
3. In the last month, how often have you felt that things were going your way?
4. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?

Tabel 6. **Søvn værdien**: Cronbach alpha: 0,75. Høj værdi udtryk for "god søvn".

1	Hvor ofte ...	er du vågnet flere gange og har haft svært ved at falde i søvn igen inden for de sidste 4 uger?
2	Hvor ofte ...	har du følt, at du ikke var udhvilet, når du vågnede inden for de sidste 4 uger?
3	Hvor ofte ...	har du følt dig træt i løbet af dagen indenfor de sidste 4 uger?

Spørgsmålene kommer/er inspireret fra Karolinska sleep questionnaire. Den oprindelige skala indeholder en del flere spørgsmål.

Tabel 7. **Energi værdien**: Cronbach alpha: 0,87. Høj værdi udtryk for god energi.

1	Hvor stor en del af tiden i de sidste 4 uger ...	har du følt dig veloplagt og fuld af liv?
2	Hvor stor en del af tiden i de sidste 4 uger ...	har du været meget nervøs?
3	Hvor stor en del af tiden i de sidste 4 uger ...	har du været så langt nede, at intet kunne opmuntre dig?
4	Hvor stor en del af tiden i de sidste 4 uger ...	har du følt dig rolig og afslappet?
5	Hvor stor en del af tiden i de sidste 4 uger ...	har du været fuld af energi?
6	Hvor stor en del af tiden i de sidste 4 uger ...	har du følt dig trist til mode?
7	Hvor stor en del af tiden i de sidste 4 uger ...	har du følt dig udslidt?

Spørgsmålene til velbefindende kommer fra undergrupperne 'vitalitet' og 'mental health' fra spørgeskemaet sf-36.(*Bjorner JB et al. The Danish SF-36 Health Survey: translation and preliminary validity studies) se* http://www.spoergeskemaer.dk/ Ikke alle spørgsmålene er inkluderet. Se også The Mental Health Inventory (MHI-5)

Tabel 8. **Stress værdien**: Cronbach alpha: 0,87. Høj værdi udtryk for lidt stress.

1	Hvor stor en del af tiden i de sidste 2 uger ... har du følt dig trist til mode, ked af det?
2	Hvor stor en del af tiden i de sidste 2 uger ... har du manglet interesse for dine daglige gøremål?
3	Hvor stor en del af tiden i de sidste 2 uger ... har du følt, at du manglede energi og kræfter?
4	Hvor stor en del af tiden i de sidste 2 uger ... har du haft mindre selvtillid?
5	Hvor stor en del af tiden i de sidste 2 uger ... har du haft dårlig samvittighed eller skyldfølelse?
6	Hvor stor en del af tiden i de sidste 2 uger ... har du følt, at livet ikke var værd at leve?
7	Hvor stor en del af tiden i de sidste 2 uger ... har du haft besvær med at koncentrere dig, fx om at læse avis eller følge med i fjernsyn?
8	Hvor stor en del af tiden i de sidste 2 uger ... har du følt dig rastløs?
9	Hvor stor en del af tiden i de sidste 2 uger ... har du følt dig stille eller fåmælt?
10	Hvor stor en del af tiden i de sidste 2 uger ... har du haft besvær med at sove om natten?
11	Hvor stor en del af tiden i de sidste 2 uger ... har du haft nedsat appetit?
12	Hvor stor en del af tiden i de sidste 2 uger ... har du haft øget appetit?

Disse spørgsmål kommer fra MDI (the Major Depression Inventory). I modsætning til de tidligere omtalte skalaer er scoringen af MDI ikke 'lige ud af landevejen'. Der er 3 scoringsmetoder. MDI bygger på ICD-10 (europæisk sygdoms klassifikation) og DSM-IV (amerikansk mentalt helbred klassifikation) definition af depression. Der er en scoringsmetode for hver af disse klassifikationer og en scoringsmetode som scorer skalaen mellem 0 og 50.

Tabel 9. **Nervøsitets værdien**: Cronbach alpha: 0,79. Høj værdi så er man ikke nervøs.

1	I de sidste 4 uger, hvor meget har du været generet af ... at du pludselig bliver bange uden grund?
2	I de sidste 4 uger, hvor meget har du været generet af ... nervøsitet eller indre uro?
3	I de sidste 4 uger, hvor meget har du været generet af ... anfald af rædsel eller panik?
4	I de sidste 4 uger, hvor meget har du været generet af ... at bekymre dig for meget?

Anxiety scale kilde ukendt.

Placering af de 36 brancher i kortet

Datamaterialet omfatter respondenter fra 36 brancher. For hver branche er der udregnet et gennemsnit af respondenternes koordinater. Dermed bliver alle brancher (store som små) placeret i en af de 4 kvadranter, som et massemidtpunkt. Se nedenstående tabel

Tabel 10. Branchernes placering (som massemidtpunkt) i de 4 kvadranter.

Factor1	Facrot2	kvadrant	beskrivelse	Brancher, tallene angiver branchekoder
factor1=>0	factor2=>0	Kvad=1	Respondenter oplever beskeden lederstøtte og stort arbejdspres	20 Nærings- og nydelsesmidler 14 Træ og møbler 10 Metal og maskiner 01 Anlægsarbejde 19 Slagterier 15 Film, presse og bøger 36 Universiteter og forskning 33 Hospitaler

factor1<0	factor2=>0	Kvad=2	Der opleves høj støtte fra ledelsen og samtidigt stort arbejdspres	06 Elektronik 28 Restauranter og barer 05 Engros 34 Læger, tandlæger og dyrlæger 16 IT og telekommunikation 17 Kontor 09 Kemi og medicin
factor1<0	factor2<0	Kvad=3	Der opleves høj støtte fra ledelsen og behersket arbejdspres	22 Religiøse institutioner og begravelsesvæsen 18 Landbrug, skovbrug og fiskeri 37 Uoplyst 31 Daginstitutioner 04 Butikker 02 Opførelse og nedrivning af byggeri 07 Energi og råstoffer 25 Hotel og camping 32 Døgninstitutioner og hjemmepleje 13 Transportmidler
factor1=>0	factor2<0	Kvad=4	Respondenter oplever beskeden lederstøtte og samtidig et beskedent arbejdspres	23 Vand, kloak og affald 27 Rengøring 26 Kultur og sport 30 Transport af passagerer 12 Tekstil og papir 35 Undervisning 03 Færdiggørelse af byggeri 08 Installation og reparation af maskiner og udstyr 21 Politi, beredskab og fængsler 29 Transport af gods 11 Plast, glas og beton

Man kan eksempelvis notere, at slagterier er placeret i kvadrant nr. 1, der angiver højt tempo og beskeden lederstøtte.

Placering af værdierne i kortet.

Hver værdi-variabel/skala inddeles i 3 grupper. Der inddeles efter den tredjedel der svarer højt på værdien, og den tredjedel der svarer lavt på værdien. Der beregnes så et gennemsnit (massemidtpunkt) af de respondenter, som har lav værdi af variablen og et gennemsnit af de respondenter som har en høj værdi af variablen. Der bliver dermed placeret to værdipunkter i kortet pr. værdivariabel. I nedenstående tabel er vist i hvilket kvadrant værdierne vil blive placeret i.

Tabel 11. Placering af værdier i værdikortet.

Factor1	Facrot2	kvadrant	beskrivelse	Antal respondenter
factor1=>0	factor2=>0	Kvad=1	Respondenter oplever beskeden lederstøtte og stort arbejdspres	Lav kontrol "med livet" Dårlig søvn Dårlig energi Meget stress Mange nerver
factor1<0	factor2=>0	Kvad=2	Der opleves høj støtte fra ledelsen og samtidigt stort arbejdspres	
factor1<0	factor2<0	Kvad=3	Der opleves høj støtte fra ledelsen og behersket arbejdspres	God kontrol "med livet" God søvn God energi Lidt stress Få nerver
factor1=>0	factor2<0	Kvad=4	Respondenter oplever beskeden lederstøtte og samtidig et beskedent arbejdspres	

Engagement værdien er ikke placeret i ovenstående skema. (stort engagement findes til højre i koordinatsystemet lille engagement til venstre).

Ikke overraskende er kvadrant 1 "det dårlige" område og kvadrant 3 "det gode" område. Der er ikke taget hensyn til hvor stor spredningen er blandt de punkter, der danner massemidtpunktet.

Konklusion.

Brugen af et begrænset antal spørgsmål til at danne et to dimensionalt kort, som man kan placere både respondenter og værdier i, synes at give et fornuftigt overblik over data. De mere "grundlæggende" faktorer (ledelse og opgaveomfang) kan delvis forklare de øvrige værdier.

De 12 spørgsmål, der danner de to faktorer, vil relativt let kunne ind koopereres i andre undersøgelser. Og dermed vil andre undersøgelser, også kunne placeres i kortet.

Det der efterfølgende bør undersøges er:

- Hvor stabilt er mønstret når man nedbryder på brancher og funktioner indenfor brancher
- Hvor stabilt er mønstret over tid. Bliver faktoranalyser "ens" i de tre undersøgelsesrunder
- Hvilke konsekvenser får en vægtning

Hvad betyder den ophørte forskerbeskyttelse ved sammenligninger over tid?

Peter Linde, DST Survey, Danmarks Statistik

Indtil 1. april 2014 var 800.000 danskere udelukket fra at deltage i frivillige statistiske undersøgelser. Det påvirkede både opnåelsen og gav derudover alle undersøgelser en ekstra bias fordi udsnittet var skævt. Det medførte af survey i Danmark set i europæisk sammenhæng samlet var blandt de dårligste mht. opnåelse og bias. Efter 1. april havde danske survey ikke mere denne udfordring, men kvalitetsforbedringen udfordre sammenligninger tilbage til i tiden, hvor der var forskerbeskyttelse. Metodisk en kvalitetsmæssig positiv udfordring, men alligevel en udfordring, som dette indlæg vil belyse.

Fra omkring 2001 og indtil 31.marts 2014 kunne man blive forhåndsfritaget for at deltage i frivillige statistiske undersøgelser. 800.000 havde 31. marts 2014 som nævnt denne status i CPR, svarende til 13% af befolkningen automatisk var bortfald ved gennerelle undersøgelser, fx borge- og tilfredshedsundersøgelser. Og noget tilsvarende i mere konkrete undersøgelser som fx ældreundersøgelser om ønsker til fremtidig bolig, hjælpen til ledige eller studieundersøgelser. Alle undersøgelser, der brugte CPR registerets navne- og adresseoplysninger som grundlag for at kontakte folk, var omfattet af loven om den såkaldte forskerbeskyttelse. Loven blev ophævet 1. april 2014 af et enigt Folketing. Undersøgelserne er fortsat frivillige – der er ikke ændret herved. Men man kan kontakte folk og spørge om de vil deltage. Danskerne bliver i gennemsnit kontakte mellem hvert femte og tiende år om en undersøgelse udvalgt fra CPR. Da man efter 1. april igen kunne kontakte de 800.000 om de ville deltage i frivillige undersøgelser svarede ca. 45% ja tak hertil. Det var ca. 15% mindre end den resterende befolkning, hvor ca. 60% svarede ja tak. Knap halvdelen af de 13% med forskerbeskyttelse ville gerne deltage, hvilket svarede til en forbedring på ca. 6% (45% af 13%) af den reelle opnåelse. De undersøgelser, der havde udelukket de forskerbeskyttede fra bortfaldsberegningen, og efter 1. april ikke mere havde denne begrænsning, oplevede et formel fald i opnåelsen fra 60% til ca. 58%.

De 800.000 svarende som nævnt til 13% af befolkningen. De havde på en række punkter en skæv fordeling. Fx var 20-25% af de 20-39 årige og ca. 20% af de arbejdsløse forskerbeskyttede. De forskerbeskyttede havde også en tendens til at svare anderledes end tilsvarende ikke forskerbeskyttede. Fx lægger forskerbeskyttede i mindre omfang penge til side, mener den økonomiske situation om et år er lidt bedre og tænker lidt ofte på risikoen for kriminalitet. Dette belyses yderligere i indlægget.

The effects of parental leave policy on the labour market outcomes of mothers and fathers

Sarah Kildahl Nico Nielsen, Rambøll Management Consulting

Introduction

Economic theory offers several explanations for why parental leave can be an important policy instrument, influencing labour market outcomes of parents. Becker (1991) introduces a family economics model, in which families achieve optimisation through specialisation. The mother and the father build human capital in either the household sector or the labour market sector and will have a higher return in the sector that they are the most specialised in. Parental leave with compensation gives incentives to specialise more in household human capital, as it increases the return to staying at home, potentially leading to lower labour market participation and wages. Spence (1973) introduced the concept of job market signalling, which can also be used to explain why take-up of parental leave might affect labour market outcomes. In a labour market with asymmetric information, the take-up of leave might be seen by employers as a signal that the employee will prioritise family over career, thereby being a less productive employee in the future, potentially leading to higher unemployment of the parents and lower wages.

This paper studies a reform of the Danish parental leave law made in 2002 and the effect of the reform on unemployment and wages. The studied reform extended shared parental leave from 10 weeks to a total of 32 weeks, at the same time as the leave reserved for the father was cut down from 4 weeks to 2 weeks.[1] Furthermore, the reform abolished childcare leave, which was a leave lasting up to 26 weeks, that could be taken by parents until their child is 9 years old. Thus the reform changed a number of mechanisms. The 2002 reform is already studied in the papers Nielsen (2009) and Beuchert, Humlum, & Vejlin (2014). Nielsen (2009) uses a difference-in-difference method, to get causal estimates, and finds that economic incentives are important for

[1] LOV nr 141 af 25/03/2002 (Civilstyrelsen, 2015)

the take-up of leave of publically employed parents; privately employed parents are not a part of the analysis. Beuchert, Humlum, & Vejlin (2014) use Regression Discontinuity Design (RDD) to get causal estimates, and find no significant effects of the reform on child and maternal health outcomes. This paper uses the same method to get causal estimates as the Beuchert, Humlum, & Vejlin (2014) paper. Another paper using RDD is Rasmussen (2010), which studies a reform of Danish parental leave law in 1984 and finds that the increase in shared parental leave and paternity leave (leave reserved for the father) in 1984 had a positive effect on mothers' working experience (labour market participation) and wages in the short run, but not the medium and long run. This paper also uses the RDD method, and finds overall only very weak evidence of an effect on mothers' labour market outcomes, but some indications of a positive effect on fathers' outcomes.

In the next section the data and methodology will be presented. After that the results will be presented and in the final section conclusions will be discussed.

Data and methodology

This section presents the methodology and the data used in the empirical analysis.

The data used is high quality Danish register data. To determine the relevant sample and to identify the causal treatment effects, the date of birth of the child is a central variable. The date of birth of children born right before and right after the studied reform is connected with data on the parents. In the Regression Discontinuity (RD) regression two variables are included, which are defined using the birth date of the children. These variables are called the treatment variable and the assignment variable. The treatment variable is a dummy variable equal to one, if the child of the parent is born on or after the date where the reform was implemented, and equal to zero if born before. The assignment variable is the birth date of the child, normalised at the date when the reform was implemented.

To analyse the mechanisms, which lead to changes in labour market outcomes, an analysis of the change in take-up of parental leave and childcare leave, is carried out.

Data on benefits received from the DREAM database is used to identify the take-up of parental leave and childcare leave for each parent on a weekly basis. An issue with this data is that it cannot be identified, which specific child/children the leave taken relates to. Therefore, when measuring the take-up of parental leave, which relates to the relevant childbirths, it is assumed that all the leave taken within the period, where the law grants the right to compensated parental leave is parental leave taken in connection with that specific child birth.[2,3] Childcare leave is only counted, if taken within the first 1.5 years (78 weeks) after the childbirth. The labour market outcomes analysed are unemployment degree per person, and wages[4] in each year in a 9 year period after childbirth.

For each of these outcomes the analysis is made separately for fathers and mothers.[5] The used model is presented below.

Model 1: $Y_{it} = \beta_0 + \tau_1 * W_i + \beta_1 * X_i + \beta_2 * Y_{i0} + \varepsilon$

Y_{it} is the outcome (unemployment, wages) in year t

β_0 is a constant

W_i is the treatment dummy variable

X_i is the assignment variable

Y_{i0} is a lagged outcome variable, measured at year 0, year 0 is here defined as approximately 2 years before the relevant births (1999), i.e. before conception

---

[2] Before the 2002 reform mothers could take compensated parental leave from 4 weeks before the birth until 24 weeks after the birth and the father could take compensated leave from the birth of the child until 26 weeks after the birth. After the 2002 reform mothers have the right to compensated parental leave at the full unemployment benefits rate from 4 weeks before the birth until 46 weeks after the birth, if she takes the full 32 weeks of shared parental leave. Since the father can take leave either at the same time or in extension of the mother's leave, he is also assumed to take his leave within the first 46 weeks after childbirth.

[3] A robustness check has been made, excluding parents, which have had one more child within one year of the relevant childbirth from the sample. The results of the analysis are very similar.

[4] The logarithm of the wages is used in the regressions, to avoid giving too much weight to very high or very low observations. Zero wages are included by adding 1 to the wage, before taking the logarithm.

[5] This allows all the parameters in the function to be different between fathers and mothers. Analysing the differences in effects for different subgroups of the parents, e.g. highly educated vs. less educated is not within the scope of the analysis in this paper.

Model 1 is the most basic RDD model.[6] The main parameter of interest is τ_1, which is the estimated Local Average Treatment Effect (LATE). Covariates in the form of Y_{i0} are included, so as to reduce the sampling variability in the estimator (Lee & Lemieux, 2010). Y_{i0} is measured before conception and thus should be uncorrelated with treatment. Furthermore, as it is a lagged value of the outcome value and some persistence is expected in the outcome variables, it improves the explanatory power of the model significantly.

In order to obtain causal estimates of the effect of the reform analysed on labour market outcomes RDD[7] is used. The idea of the RDD is to compare parents to children born right before the implementation of the reform (the control group) with parents to children born right after the implementation of the reform (the treatment group). The disadvantage of using RDD is the low external validity, and that we do not get an estimate for the effect of take-up of leave on labour market outcomes, but simply an estimate for the overall effect of the studied reform. For the RDD method to work, it is necessary that there is a discontinuity in data around the treatment. For the studied reform parents to children born between the 1st of January 2002 and the 26th of March could freely choose, which parental leave policy they wanted to be part of, making this discontinuity a little fuzzy. Beuchert et al. (2014) argues that the assumption holds, based on the empirics, which show that the take-up of leave takes a sharp jump at the 1st of January 2002. The discontinuity is therefore assumed to be sufficient.

The main identifying assumption of RDD is that parents cannot perfectly sort around the treatment date. Under this assumption the LATE can be identified, giving causal estimates of the effect of the parental leave reforms on labour market outcomes for the sample analysed. Perfect sorting around the treatment date of the reform, is very unlikely. The reform was announced quite shortly before it was implemented, making

[6] More flexible model specifications, such as including a squared assignment variable or an interaction term for the treatment dummy and the assignment variable have been tested and did not significantly improve the results.

[7] The focus here is on sharp RDD and not fuzzy RDD. For more about the RDD method see Imbens & Lemieux (2008).

it impossible for parents to plan the conception such that their child should be born under the old or the new policy. Caesarean sections or drug induced labour give the possibility of some parents planning the exact time of birth of their child, but both are usually only performed given certain health issues of the mother or child and most parents in Denmark are therefore not free to choose the birth date of their child.[8]

To get causal estimates using RDD the sample needs to be limited to parents of children born right before and right after the reform analysed. The size of the sample is based on the trade-off between bias and variance. The larger the sample the more variance is achieved and the easier it will be to get precise and/or significant estimates, but a larger sample also increases the probability that the control group and the treatment group will be significantly different with regards to observable background characteristics. Based on this trade-off the main sample is defined as parents to children born between the 1st of November 2001 and the 1st of March 2002. The final sample consists of 17,845 fathers and 18,008 mothers. Robustness analyses of the sample size have been made, looking at smaller or larger intervals of time around the time of birth of the child.[9] A method to test for sorting around the treatment date is to test for differences in observable background characteristics between the control and the treatment groups.

Table 1 shows sample statistics for the control and treatment group, for men and women respectively. The table presents means for the control and treatment group and test for a significant difference in means. A significant difference in means indicates that the identifying assumption does not hold. As an alternative test of difference between the two groups, a RD regression is made, for each covariate, using the covariate variable as the dependent variable. A significant coefficient for the treatment dummy means that the covariate is discontinuous at the treatment date, indicating sorting around the treatment date.

[8] Only about 7 percent of all births were planned as caesarean sections before the birth in 1998 (Sundhedsstyrelsen, 2005)

[9] The results of the robustness analyses are summarized in the conclusions. To access the output from the analyses, please contact the author.

TABLE 1: SAMPLE STATISTICS FOR THE CONTROL AND TREATMENT GROUP, 2002 REFORM

	Control		Treatment		Diff. means	RD	
	Mean	*Std. dev.*	*Mean*	*Std. dev.*	*P-value*	*Coeff.*	*Std. err.*
Men							
Work experience (years)	7.76	0.6	7.79	0.05	0.76	0.20	0.16
Unemployment degree (percent)	4.27	0.14	4.43	0.15	0.43	-0.36	0.42
Net wage (1.000 kr.)	216.37	1.65	217.37	1.55	0.39	8.71*	4.59
Age (years)	30.49	0.06	30.35	0.06	0.69	0.16	0.17
Number of children	0.68	0.01	0.68	0.01	0.92	0.03	0.03
Couple (percent)	76.22	0.46	76.34	0.45	0.85	1.90	1.29
High school education (percent)	74.70	0.47	76.36	0.45	0.01	1.49	1.31
Women							
Work experience (years)	5.09	0.04	5.13	0.04	0.57	0.16	0.13
Unemployment degree (percent)	6.14	0.17	6.20	0.16	0.81	0.18	0.47
Net wage (1.000 kr.)	136.77	1.15	137.41	1.08	0.68	2.44	3.19
Age (years)	27.87	0.05	27.80	0.05	0.36	0.30**	0.15
Number of children	0.78	0.01	0.77	0.01	0.41	0.02	0.03
Couple (percent)	77.32	0.45	77.64	0.44	0.60	2.27*	1.26
High school education (percent)	74.14	0.47	75.22	0.45	0.10	3.25**	1.31

Source: *Statistics Denmark and calculations made by the author.*
Note: *Covariates statistics are from 1999, except number of children which is from primo 2000. The control group is parents to children born from November 1st to December 31st 2000 and the treatment group consists of children born from January 1st to February 28th 2001. The test for difference in means is a standard t-test. The RD estimates reported are the coefficients for the treatment dummy when using the covariate as dependent variable. The assignment variable is also included in the specification but not shown here. The standard errors are robust. * significant at 10 percent level, ** significant at 5 percent level, *** significant at 1 percent level*

At a 5 percent significance level, the standard t tests show no indication of difference in means between the two groups, except for the share with a high school education. The RD regression finds a few significant estimates, but overall the table supports the 2nd identifying assumption, i.e. that the expected outcomes of the control and treatment groups are continuous for parents to children born right before and right after the implementation of the reform. There is some indication that the mothers in the treatment group have a higher education than the mothers in the control group.[10]

[10] A higher education might bias the results towards a positive estimate for the reform for mothers, but this is only found in one year and the results do therefore not seem to be biased.

Results

In this section the results will be presented and analysed to find the effect of the reform on gender equality in the analysed labour market outcomes.

Take-up of parental and childcare leave

Table 2 shows standard t tests for difference in means between the control and treatment group, as well as estimated RD coefficients with take-up of leave as the dependent variable. The methods used are the same as used to test for differences in Table 1.

Table 2: Take-up of parental leave before and after reforms (weeks), 2002 reform

	Control		**Treatment**		**Diff. means**	**RD**	
	Mean	*Std. dev.*	*Mean*	*Std. dev.*	*P-value*	*Coeff.*	*Std. err.*
Parental leave							
Father	2.19	0.03	2.33	0.04	0.00	-0.06	0.09
Mother	23.10	0.13	34.52	0.19	0.00	11.52***	0.47
Childcare leave							
Father	0.67	0.05	0.29	0.04	0.00	-0.49***	0.12
Mother	12.59	0.18	3.89	0.13	0.00	-8.22***	0.44
Total leave							
Father	2.86	0.06	2.63	0.05	0.00	-0.55***	0.15
Mother	35.70	0.24	38.41	0.22	0.00	3.29***	0.67

Source: *Statistics Denmark and calculations made by the author.*
Note: *The test for difference in means is a standard t-test. The RD coefficients are for the treatment dummy when using the covariate as dependent variable. The assignment variable is also included in the specification but not shown here. The standard error is robust. * significant at 10 percent level, ** significant at 5 percent level, *** significant at 1 percent level.*

Mothers increase their take-up of parental leave significantly, whereas the standard t-test and the RD regression lead to different conclusions about the fathers' take-up of parental leave. Looking at differences in means, fathers increase their take-up of parental leave with 0.14 weeks or approximately 1 day, while mothers increase their take-up of parental leave with 11.42 weeks or approximately 80 days. According to the standard t-test fathers take significantly more leave after the reform, but the RD

regression finds no significant effect. Both results would be consistent with theory, as there are opposite effects of expanding the shared parental leave and shortening the paternal leave. Thus the t-test indicates that the effect of expanding the shared parental leave was greater than the effect of shortening the paternity leave, while the RD regression indicates that the effects of the two changes cancel each other out.

Table 2 further shows that both fathers' and mothers' take-up of childcare leave is lower for the treatment group than for the control group. The reason that take-up of childcare leave is not zero for the treatment group could be that the childcare leave registered is in fact saved leave from a previous child or that not all parents in the treatment group chose to be covered by the new policy. The overall effect is that the total leave for fathers decreases significantly, while the total leave for mothers increases significantly.

Unemployment

In Table 3 it is seen that the constant and the lagged outcome variable are significant determinants of the unemployment for fathers and mothers. Table 3 shows some indications of a significant treatment effect of the 2002 parental leave reform on unemployment of fathers in the period 2003-2011. The effect in 2003 is almost significant and there is a weakly significantly negative effect on unemployment of fathers in 2004, this could indicate some negative effects on unemployment of fathers in the short run. After 2004 the estimates are very insignificant. A negative effect on unemployment of fathers indicates that they increase their labour market participation, which is what the theory predicts, when the fathers decrease their take-up of leave.

Table 2 shows that when considering childcare leave, fathers clearly decreased their take-up of leave, the negative effects in the beginning are consistent with theory, given that fathers used to take their childcare leave, in these first years after childbirth. The effects might only be temporary because the fathers that take more leave, i.e. fathers to children born right before the 2002 parental leave reform, after some time manage to catch up with regards to human capital accumulation or signalling high productivity. Overall the evidence for an effect on unemployment of fathers is somewhat weak.

TABLE 3: REGRESSION DISCONTINUITY ESTIMATES OF THE EFFECT OF THE REFORM ON UNEMPLOYMENT IN 1999-2007 (PCT. POINTS)

	Unemp. in 2003	Unemp. in 2004	Unemp. in 2005	Unemp. in 2006	Unemp. in 2007	Unemp. in 2008	Unemp. in 2009	Unemp. in 2010	Unemp. in 2011
Fathers									
Treatment	-0.6578	-0.7954*	-0.1622	0.2471	0.3581	0.1660	-0.3456	-0.1316	-0.3601
	(0.4703)	(0.4625)	(0.4236)	(0.3710)	(0.3342)	(0.2208)	(0.3599)	(0.3668)	(0.3345)
Date of birth	0.0069	0.0075	-0.0009	-0.0060	-0.0078*	-0.0057*	-0.0019	-0.0001	0.0030
	(0.0067)	(0.0068)	(0.0061)	(0.0530)	(0.0047)	(0.0032)	(0.0052)	(0.0052)	(0.0047)
Unemployment	0.2657***	0.2425***	0.2002***	0.1449***	0.1355***	0.0690***	0.0962**'	0.0893***	0.0835***
in 1999	(0.0150)	(0.0153)	(0.0136)	(0.0119)	(0.0119)	(0.0072)	(0.0099)	(0.0101)	(0.0092)
Constant	4.3421***	4.4828***	3.3010***	2.3887***	1.6994***	1.0813***	3.1784***	3.0820***	2.6213***
	(0.2744)	(0.2736)	(0.2460)	(0.2101)	(0.1909)	(0.1265)	(0.2095)	(0.2115)	(0.1942)
Obs.	17,845	17,845	17,845	17,845	17,845	17,845	17,845	17,845	17,845
R^2	0.0532	0.0448	0.0380	0.0263	0.0294	0.0168	0.0125	0.0100	0.0108
Mothers									
Treatment	2.1767***	-0.3526	0.1467	-0.5515	-0.7854*	-0.3079	-0.3619	0.3345	-0.0233
	(0.6182)	(0.5837)	(0.5376)	(0.4838)	(0.4293)	(0.2537)	(0.3139)	(0.3161)	(0.3509)
Date of birth	-0.0096	0.0657	-0.0067	0.0288	0.0079	0.0014	0.0039	-0.0034	-0.0017
	(0.0087)	(0.0082)	(0.0076)	(0.0678)	(0.0061)	(0.0036)	(0.0045)	(0.0046)	(0.0051)
Unemployment	0.2590***	0.2138***	0.1748***	0.1104***	0.1101***	0.0505***	0.0574***	0.0437***	0.0519***
in 1999	(0.0130)	(0.0125)	(0.0117)	(0.0098)	(0.0097)	(0.0061)	(0.0070)	(0.0067)	(0.0075)
Constant	7.6715***	8.0289***	6.5165***	5.8105***	4.5303***	1.9295***	2.7550***	2.3069***	2.8045***
	(0.3482)	(0.3369)	(0.3080)	(0.2767)	(0.2465)	(0.1444)	(0.1819)	(0.1819)	(0.2001)
Obs.	18,008	18,008	18,008	18,008	18,008	18,008	18,008	18,008	18,008
R^2	0.0393	0.0297	0.0234	0.0121	0.0154	0.0090	0.0071	0.0042	0.0051

Source: *Statistics Denmark and calculations made by the author.*

Note: *Robust standard errors in parentheses. * significant at 10 percent level, ** significant at 5 percent level, *** significant at 1 percent level.*

TABLE 4: REGRESSION DISCONTINUITY ESTIMATES OF THE EFFECT OF THE REFORM ON WAGES IN 1999-2007 (LOG-UNITS)

	Wage in 2003	Wage in 2004	Wage in 2005	Wage in 2006	Wage in 2007	Wage in 2008	Wage in 2009	Wage in 2010	Wage in 2011
Fathers									
Treatment	0.1867*	0.2093*	0.1134	0.0382	-0.0138	-0.0442	0.0065	-0.0043	0.0054
	(0.1119)	(0.1149)	(0.1165)	(0.1174)	(0.1208)	(0.1229)	(0.1276)	(0.1327)	(0.1346)
Date of	0.0010	-0.0015	-0.0011	-0.0004	0.0007	0.0005	-0.0003	0.0008	0.0005
birth	(0.0016)	(0.0016)	(0.0017)	(0.0017)	(0.0017)	(0.0017)	(0.0018)	(0.0019)	(0.0019)
Wage in	0.5826***	0.5493***	0.5095***	0.4768***	0.4537***	0.4470***	0.4672***	0.4736***	0.4745***
1999	(0.0101)	(0.0103)	(0.0105)	(0.0105)	(0.0106)	(0.0107)	(0.0108)	(0.0109)	(0.1019)
Constant	4.3950***	4.7797***	5.3483***	5.8543***	6.1914***	6.2748***	5.8239***	5.6205***	5.6166***
	(0.1355)	(0.1383)	(0.1402)	(0.1407)	(0.1425)	(0.1433)	(0.1449)	(0.1471)	(0.1479)
Obs.	17,845	17,845	17,845	17,845	17,845	17,845	17,845	17,845	17,845
R^2	0.2935	0.2577	0.2248	0.1838	0.1807	0.1694	0.1702	0.1618	0.1606
Mothers									
Treatment	0.1229	0.0038	-0.0383	0.0301	-0.0970	0.0210	0.1126	0.0331	-0.0074
	(0.1305)	(0.1323)	(0.1290)	(0.1266)	(0.1232)	(0.1214)	(0.1260)	(0.1310)	(0.1316)
Date of	-0.0015	0.0020	0.0025	0.0006	0.0016	0.0001	-0.0011	-0.0002	-0.0001
birth	(0.0019)	(0.0018)	(0.0018)	(0.0018)	(0.0017)	(0.0017)	(0.0018)	(0.0019)	(0.0019)
Wage in	0.5564***	0.5091***	0.4899***	0.4481***	0.4079***	0.3855***	0.3991***	0.4084***	0.4158***
1999	(0.0088)	(0.0091)	(0.0092)	(0.0093)	(0.0093)	(0.0093)	(0.0094)	(0.0096)	(0.0096)
Constant	3.9048***	4.6421***	5.0567***	5.7187***	6.4659***	6.7892***	6.5590***	6.4021***	6.3243***
	(0.1210)	(0.1240)	(0.1248)	(0.1247)	(0.1239)	(0.1238)	(0.1267)	(0.1290)	(0.1293)
Obs.	18,008	18,008	18,008	18,008	18,008	18,008	18,008	18,008	18,008
R^2	0.2443	0.2116	0.2059	0.1838	0.1634	0.1507	0.1519	0.1489	0.1505

Source: *Statistics Denmark and calculations made by the author.*

Note: *Robust standard errors in parentheses. * significant at 10 percent level, ** significant at 5 percent level, *** significant at 1 percent level.*

Table 3 shows some indications of a significant effect of the 2002 parental leave reform on unemployment of mothers. Mainly there seems to be a positive effect on unemployment in the short term, i.e. a negative effect on labour market participation. This is likely to be explained by the steep increase in the mothers' take-up of parental leave seen in

Table 2, but it seems to only have a positive effect on unemployment of mothers in 2003, the year after the maternity leave was taken. In the longer run negative estimates are seen for the effect in most years, but the effect is only significant in 2007. A positive effect in the short run and a negative effect on unemployment in the longer run are consistent with theory, if the reduction in take-up of childcare leave happens in later years. In this case the 2002 reform would make leave-taking increase right after childbirth, leading to the positive bump in unemployment seen in 2003. But in later years after the childbirth the reform would lead to a decrease in leave-taking and with that more negative effects on unemployment. If women take childcare leave later than 1.5 year after childbirth, it should be noted that it would not be included in the measure for childcare leave described in

Table 2, and the reduction in childcare leave is therefore possibly bigger than what is described.

Wages

Table 4 shows the RD regression estimates for wages of fathers and mothers in the period 2003-2011. For the earnings regressions for fathers and mothers it is seen that the constant and the lagged outcome variable have a significant effect on wages in the analysed years, and that the explanation power of the model (measured by the R^2) is higher than the model for unemployment. The estimated effects for wages of fathers support the vague results found for unemployment, i.e. that there is a positive short-term effect. Table 4 show a significant positive effect of the 2002 parental leave reform on wages of fathers in the years 2003 and 2004. This is consistent with theory, since total take-up of leave including childcare leave is reduced after the policy reform. In later years the difference between the control and treatment group of fathers

disappears. As for unemployment this is likely to be because the fathers who took more leave over time manage to catch up with the ones who did not. No indications of a significant treatment effect for mothers wages of the 2002 parental leave reform is found in Table 4. The effects on unemployment therefore do not carry over to changes in wages for mothers. This is likely because the opposing effects of extending parental leave and abolishing childcare leave, cancel each other out.

Conclusions

Overall there is some evidence that the 2002 parental leave reform had positive effects on labour market participation and wages of fathers, which likely is due to the decrease in the total take-up of leave that was seen in

Table 2. The result for unemployment is fairly stable for different sample definitions (different time intervals around the treatment date), while the positive treatment effect for wages is less robust. For mothers there is some evidence that unemployment is affected, with a positive effect on unemployment i.e. a negative effect on labour market participation in the very short run, but negative effects on unemployment in later years. This is likely to be due to the structure of the 2002 parental leave law, as discussed. Looking at different sample sizes (different time intervals around the treatment date) there is some indication that the effect on unemployment is more negative than the main analysis indicates. For wages, it is seen that the significant effect on unemployment does not carry over to wages. The estimate for the positive effect on unemployment in 2003 is larger than the estimate for the negative effect on unemployment in 2007, but the difference between the two estimates is not significant. It is therefore likely that the effect of expanding shared parental leave was cancelled out by the effect of abolishing childcare leave for mothers' wages.

Reference List

Becker, G. (1991). *Treatise on the Family*. Cambridge: Harvard University Press.

Beuchert, L. V., Humlum, M. K., & Vejlin, R. (2014). The Length of Maternity Leave and Family Health. *IZA Discussion Papers, 8206*.

Civilstyrelsen (2015). Retsinformation. Retrieved 29-9-2015, from https://www.retsinformation.dk/

Imbens, G. W. & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142,* 615-635.

Nielsen, H. S. (2009). Causes and Consequences of a Father's Child Leave: Evidence from a Reform of Leave Schemes. *IZA Discussion Papers.*

Rasmussen, A. W. (2010). Increasing the length of parents' birth-related leave: The effect on children's long-term educational outcomes. *Labour Economics, 17,* 91-100.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics, 87,* 355-374.

Sundhedsstyrelsen (2005). *Kejsersnit på moders ønske - en medicinsk teknologivurdering.*

Forecasting macroeconomic labour market flows: What can we learn from micro level analysis?

Ralf A. Wilke*

December 2015

Abstract

Forecasting labour market flows is important for budgeting and decision making in government departments and public administration. Macroeconomic forecasts are normally obtained from time series data. In this paper we follow another approach that uses individual level statistical analysis to predict the number of exits out of unemployment insurance claims. We present a comparative study of econometric, actuarial and statistical methodologies that base on different data structures. The results with records of the German unemployment insurance suggest that prediction based on individual level statistical duration analysis constitutes an interesting alternative to aggregate data based forecasting. In particular forecasts of up to 6 months ahead are surprisingly precise. Econometric forecasts based on time series data are found to have a larger mean squared error due to the small number of observations they use.

Keywords: unemployment insurance, in-sample forecasting, duration analysis

*Copenhagen Business School, Department of Economics, Porcelænshaven 16A, 2000 Frederiksberg, Denmark, E–mail: rw.eco@cbs.dk

Early Labour Market Disruption:
Effect of Young Adult Childbearing on the Women's Labour Market Outcome

Philip Rosenbaum[*]

Abstract: Work interruptions related to childbirths are expected to affect mothers' wages directly through changes in their human capital formation. This effect is proposed to be exceptionally strong for young adult childbearing women who are about to enter their working careers. This study investigates whether the long-term socioeconomic problems experienced by women with their first childbearing before turning 26 are a reflection of pre-existing disadvantages or are a consequence of the childbearing timing? The purpose is furthermore to observe whether a new combination of the best practices of earlier studies on the subject can serve as a better estimation method. This is done by applying a Sister First Difference estimator while using miscarriages as exogenous variation. This exact design has, to my knowledge, never been used before to estimate socio-economic effects of childbearing timing. I find no effects of young adult childbearing on the women's wages.

1 Introduction

When is the best time for a woman to get pregnant? Postponing motherhood may reduce the women's overall number of children, since fertility decreases with age. At the same time, there is a predominant belief that early childbearing has a negative impact on the women's educational attainments and diminishes their employment perspectives. Contrary to the common belief, this study finds no evidence that young adult childbearing has a persistent negative effect on women's wages.

I apply a Sister First Difference method on three different Sister-Samples. Each of the three samples is designed in order to shed light on different implications of young adult childbearing. The first Sister-Sample consists of sister-pairs of early and non-young adult childbearing sisters. This sample is assembled to replicate earlier studies and to show whether the same results can be obtained on Danish women. The result obtained on the basis of this sample was that the effect of young adult childbearing on wages is significantly negative in the short run (five years), but insignificant in the long run (ten years).

The Second Sister-Sample consists of young adult childbearing women and their non-young adult childbearing sisters, which have had an abortion at an early age. When using sisters with an early abortion as controls, the effect of early childbearing was very large, implying that the conscious choice of postponing the first childbirth through an abortion separates them from their early childbearing sisters.

The Third Sister-Sample contains women with early childbearing and their non-early childbearing sisters, who suffered a miscarriage at an early age. The effect of early childbearing

* Philip Rosenbaum, Department of Economics, Copenhagen Business School, Porcelænshaven 16a, Frederiksberg, Denmark

disappeared when applying a Sister First Difference estimator together with using control sisters who miscarried in an early age.
These results have many implications. First of all, they show that there may remain some unobserved heterogeneity after applying a Sister First Difference method on Sister-Sample 1, implying that there remain systematic differences between the women with young adult childbearing and their sisters. This indicates that results from earlier sister-studies on early childbearing may be biased. The remaining heterogeneity is addressed in Sister-Sample 3, when I use miscarrying sisters as controls.
The main result of this study is therefore that young adult childbearing has no persistent effect on women's wages. I.e. young adult childbearing women's inferior wage outcomes are not due to having a child in an early age, but rather due to pre-existing disadvantages in social- and ability factors.

2 Literature Review

Women who delay childbirth are experiencing higher wages, which there basically can be two reasons for: 1. The mommy track where childbirth leads to a lower wage rate 2. The main reason for this is the lower human capital experienced by mothers. Becker's Household Production Theory (1965) implies that the opportunity cost of working increases when getting a child and thus the effort and productivity will decrease at the workplace. (Gronau (1974), Bronar, Stephen & Groggar (1994) and Angrist & Evans (1998)). Reverse causality where early childbearing women essentially would not have performed well at the labour market even without childbearing. Especially a drop in the human capital investments in the start twenties – both as a result of disruptions in the education or at the job - are shown to have long term negative effects on the labour market outcome (Gerster et al 2014). This effect is called Scarring and refers to the poor habits developed in periods of labour market disruptions, which catalyse persistent labour market detachment and alienation. (Ellwood (1983), Gartell (2009)).

By reversing the causality on the relation of young adult motherhood and adult wages, the timing of the first childbearing can be seen as an indicator of the women's endowed human capital and not a consequence of the time and effort motherhood cost. If there is a reverse causality, then the timing of the first childbirth might be an economic indicator for the woman's productivity and her preference towards a working career.[1] I.e. their price of time is lower than for high-productive women, which is what Gronau (1974) called the shadow-prices of early childbearing.

[1] Of cause childbirth cannot be planed to the minute, but on average it is possible to time the childbirth in accordance to the women's career plan.

3 Empirical Approach

My empirical strategy is an extension of the methods originally used in the young mother empirical literature and it is specifically designed to elicit the true effects of having a child as a young adult.

There have been two main approaches designed in order to cope with the family heterogeneity and the individual unobserved heterogeneity respectively.[2] The first approach designed to account for family heterogeneity is the within-family estimates. I.e. comparing sisters where one gave birth in her youth while the other did not. The idea is that the remaining differences between the sisters' socioeconomic outcomes primarily will be due to the difference in their age at first childbirth (Geronimus & Korenman (1992) (Hoffman et al. (1993), Rosenzweig and Wolpin (1995) and Holmlund (2004)).

The second approach is to exploit exogenous variations or institutional changes in order to account for individual unobserved heterogeneity. The most relevant for this study was originally conducted by Hotz et al. (1997) & (2005), who studied teenage pregnancies, while Miller (2011) studied effects of motherhood timing on career paths, both using miscarriages as an instrument. They looked at early childbearing mothers and compared them with other women who conceived at the same age but underwent a miscarriage and therefore postponed childbearing.

I will estimate the effect of young adult childbearing on the women's yearly wage, by applying a within-family method on three different Sister-Samples of Danish women. The idea is to apply a combination of the two econometric approaches described above. The within-family approach will cope with the unobserved family heterogeneity and conditioning the control sisters – the sisters of the young adult mothers – on having had an abortion or a miscarriage in as a young adult, should work as exogenous variation ensuring a random assignment of the sisters to the control and treatment group. Furthermore, I control for the women's general health history. All together this novel method will remove the biases that otherwise could have poisoned the results.

3.1 Sister First Differences as a Mean of Removing Bias

Being able to collect information about individuals and their families allows me to organize the dataset in a panel structure. The panel consist of two sisters per family. The sisters have family invariant variables as well as family variant variables.

[2] There have been used other identification strategies, which are less relevant for this study, and arguably less precise. E.g. Matching method (Simonsen & Skipper 2006)

One way to deal with unobserved heterogeneity is by applying a sister first difference model. Its differencing transformation has a very pleasing application in this situation. I withdraw the sister values from each other:

$$y_{ij} = \gamma YM_{ij} + \beta_1 X_{1j} + \beta_2 F_j + \alpha_j + \mu_{1j} \qquad (1)$$

Where YM_{ij} is a dummy indicating young adult childbearing, X_{ij} is the family and individual variant variables - such as the woman's age, number of diagnosis, F_j is the family invariant variables – such as region of residence in adolescence, immigration status, parents' education. Let α_j be the unobserved family heterogeneity variable. Unobserved heterogeneity is the same for all members of the same family- e.g. parental involvement.[3] If α_j is ignored and it is correlated with the other explanatory variables, the OLS estimates are bound to be biased. μ_{ij} is the new idiosyncratic error term. Only the difference between the sisters will remain after withdrawing y_2 from y_1:

$$\Delta y_j = \gamma \Delta YM_j + \beta_1 \Delta X_j + \Delta \mu_j \qquad (2)$$

Equation (3) is the reduced model, where; $y_j = y_{1j} - y_{2j}, \Delta YM_j = YM_{1j} - YM_{2j}$, $\Delta X_j = X_{1j} - X_{2j}, and\ \Delta \mu_j = \mu_{1j} - \mu_{2j}$. This transformation removes all the family invariant variables - both the observable, **F**, and the unobservable, α.[4] All of the unobserved heterogeneity will be removed if it only consists of the sisters' shared environment.

3.1 Methodologically Advantages and Limitations

Using sisters may provide a good way of accounting for unobserved family background characteristics, but heterogeneity certainly also exist within families. Siblings may vary in unobservable factors. Such as their endowments or in the extent and fashion in which their parents invest in the sisters (Berhman & Taubman (1986), Ejrnæs & Portner (2004) and Black (2005)).

3.2 Sample Selection

As described, the populations in this study consist of three Sister-Samples. The control sisters in all three Sister-Samples have not had childbirth as a young adult, where the sisters in the second sample had an abortion and in the third sample had a miscarriage as a young adult. Of course random experiments are the golden standard, but have the advantages explained in more details in the following sections.

[3] Some studies have proposed that parental involvement differs between their children. Hence the parents are more involved in their first born life than in the rest of their children. This phenomenon will be discussed further later.

[4] Notice that the intercept does not appear in this model, because it also is removed through the transformation.

3.2.1 Designing the Sister-Samples

Young adult childbearing women are not randomly selected. One cannot claim that young adult childbearing is an exogenous event, implying that the event of getting a child in an early age is highly correlated with other life choices that influence socioeconomic variables. This evidently leads to selection bias problems.

To deal with the selection bias and the unobserved heterogeneity, the regression studies are performed on three different and carefully selected samples. All the samples consist of sister-pairs where one sister had a young adult childbearing - before turning 26 - and other sister did not. This is also the only restriction on Sister-Sample 1. In Sister-Sample 2 the non-early childbearing sisters are further restricted by having had an induced abortion before turning 26. In Sister-Sample 3 the non-early childbearing sisters are restricted by having had a miscarriage before turning 26. The sister-pair is placed in Sister-Sample 2 if the non-early childbearing sister had both an early abortion and a miscarriage in a young age.

Differentiating between abortions - as a conscious termination of pregnancies – and miscarriages – as a random termination of pregnancies – can have interesting suggestions.

Sister-Sample 2: The selection effect of Sister-Sample 2 is predictively ambiguous. One factor is that both sisters became pregnant as a young adult. This indicates some kind of shared lifestyle between the two of them. On the other hand, the conscious choice of getting an abortion may indicate a discrepancy in the sisters' life planning. The choice of postponing childbearing at an early age may indicate that the woman evaluates and prefers differently (e.g. education and career) than her sister with young adult childbearing. The question is which of these two opposing factors is the dominant? Or are any of these two factors even present? Is it a bigger lifestyle indicator to get pregnant in an early age than it is to choose to terminate the pregnancy?

Sister-Sample 3: The selection effect of Sister-Sample 3 is relatively one-sided, since miscarriages are not a result of a conscious decision the distribution of miscarriage occurrences can be seen as random. Because of this randomness in the pregnancy outcomes many of the selection problems disappear since the unobserved variables no longer can be systematically unevenly distributed and create unbiased estimates. But is it that simple, and is miscarriages a perfect exogenous variation?

Unsuitably, doubts on the randomness of miscarriages exist and are probably reasonable. Where the selection of Sister-Sample 2 tends to homogenise the women through their shared lifestyle at the time of pregnancy, the selection of Sister-Sample 3 may in fact do the opposite.

One could suspect that women with inferior general health and unhealthy lifestyle during the pregnancy miscarry more frequently. It might be that miscarriages are unconscious occurrences but indirectly induced by the women's behaviour, which also influences the labour market

outcomes and therefore will be problematic. It is generally perceived that people with health problems generally perform worse at the labour market and if the women who miscarry generally experience health problems, it will be difficult to distinguish whether the labour market performance is due to women's miscarriages or their poor health conditions (Smith (2009)). It is therefore of great importance to incorporate a health variable that captures the systematic health deviation between the sisters.

If health problems are properly incorporated and there exist no other systematic differences between the sisters, the Within-Family method on Sister-Sample 3 will be suitable for examining the effect of early childbearing. The exogenous variation and the sister first difference will satisfy the conditions - described above – that are needed to obtain an unbiased and consistent estimator.

4 Preliminary Conclusion

The young adult childbearing women have lower adult wages than the average women; nevertheless it is not because of their young adult childbearing, but rather due to their pre-birth backgrounds, attributes and circumstances. Hence, the main result of this study is that young adult childbearing does not have a persistent effect on women's wages.

5 Literature

ANGRIST, JOSHUA D. & EVANS, WILLIAM N. (1998): *Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size*, The American Economic Review, Vol. 88, No. 3.

BECKER, GARY S. (1965): *A Theory of the Allocation of Time*, The Economic Journal.

BEHRMAN, JERE R. & TAUBMAN, PAUL (1986): *Birth Order, Schooling, and Earnings*, Journal of Labor Economics, Vol. 4, No. 3.

BLACK S.E., P.J. DEVEREUX & K.G. SALVANES (2005): *The More the Merrier? The Effect of Family Size and Birth Order on Children's Education*, Quarterly Journal of Economics, Vol. 120, No. 2, pp 669-700

BRONAR, STEPHEN G. & GROGGER, JEFF (1994): *The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment*, The American Economic Review, Vol. 84, No. 5.

EJRNÆS, METTE & PÖRTNER, CLAUS C. (2004): *Birth Order and the Intrahousehold Allocation of Time and Education*, The Review of Economics and Statistics, Vol. 86, No. 4.

ELLWOOD, DAVID T. (1882): *Teenage Unemployment: Permanent Scars or Temporary Blemishes?*, Chapter in *The Youth Labor Market Problem: Its Nature, Causes, and Consequences.* University of Chicago Press, pp. 349-390.

GARTELL, MARIE (2009): *Unemployment and Subsequent Earnings for Swedish College Graduates: a Study of Scarring Effects*, Institute for Labour Market Policy Evaluation, Working Paper Collection

GERONIMUS, ARLINE T. & KORENMAN, SANDERS (1992): *The Socioeconomic Consequences of Teen Childbearing Reconsidered*, The Quarterly Journal of Economics, Vol. 107, No. 4.

GERSTER, METTE; EJRNÆS, METTE & KEIDING, NIELS (2014): *The Causal Effect of Educational Attainment on Completed Fertility for a Cohort of Danish Women – Does Feedback Play a Role*, Statistics in Bioscience, Vol. 6, No. 2. Pp. 204-222.

GRONAU, REUBEN (1974): *The Effect of Children on the Housewife's Value of Time*, National Bureau of Economic Research.

HOFFMAN, SAUL D.; FOSTER, E. MICHAEL & FURSTENBERG JR., FRANK F., (1993): *Reevaluating the Costs of Teenage Childbearing*, Vol. 31, Demography No. 1.

HOTZ, JOSEPH; MULLIN, CHARLES H. & SANDERS, SETH G. (1997): *Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analysing the Effects of Teenage Childbearing*, The Review of Economic Studies, 1997.

HOTZ, JOSEPH; MCELROY, SUSAN W. & SANDERS, SETH G. (2005): *Teenage Childbearing and Its Life Cycle Consequences: Exploiting a Natural Experiment*, The Journal of Human Resources, Vol. 40, No. 3, 2005.

MILLER, AMALIA R. (2011): *The Effects of Motherhood Timing on Career Path*, Journal of Population Economics, Vol. 24, pp. 1071-1100.

ROSENZWEIG, MARK R. & WOLPIN, KENNETH I. (1980): *Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment*, Econometrica, Vol. 48, No. 1.

SIMONSEN, MARIANNE & SKIPPER, LARS (2006): *An Analysis Using Matching Estimators*, Journal of Applied Econometrics, Vol. 21, pp. 919-934.

SMITH, JAMES P. (2009): *The Impact of Childhood Health on Adult Labor Market Outcomes*, The Review of Economics and Statistics, August.

Statistical methods for determining the effect of mammography screening

Søren Lophaven
Center of Epidemiology and Screening, Department of Public Health
University of Copenhagen

1 Introduction

In an overview of five randomised controlled trials from Sweden, a reduction of 29% was found in breast cancer mortality in women aged 50-69 at randomisation after a follow up of 5-13 years. Organised, population based, mammography service screening was introduced on the basis of these results in the municipality of Copenhagen in 1991, in the county of Fyn in 1993 and in the municipality of Frederiksberg in 1994, although reduced mortality in randomised controlled trials does not necessarily mean that screening also works in routine health care. In the rest of Denmark mammography screening was introduced in 2007-2008. Women aged 50-69 were invited to screening every second year.

Taking advantage of the registers of population and health, we present statistical methods for evaluating the effect of mammography screening on breast cancer mortality (Olsen et al. 2005, Njor et al. 2015 and Weedon-Fekjær et al. 2014). The results obtained when applying these methods will be briefly presented.

2 The Olsen et al. (2005) approach

The large time gap between regions with early (Copenhagen, Fyn and Frederiksberg) and late (rest of Denmark) introduction of screening makes regions with late introduction of screening a natural control group, and this can be utilised when examining whether the mammography screening programme actually reduce mortality due to breast cancer. Olsen et al. (2005) used the advantage of this natural experiment to evaluate the effect of screening in the municipality of Copenhagen. They constructed a study group

consisting of women invited to screening in the municipality of Copenhagen in the period 1 April 1991 - 1 April 2001, a historical control group consisting of women in the same age group living in Copenhagen before the start of the screening programme (in the period 1 April 1981 - 31 March 1991), a national control group consisting of women in the same age group living in Denmark outside the three screening regions in the period 1 April 1991 - 1 April 2001, and a historical national control group consisting of women in the same age group living in Denmark outside the three screening regions before start of the screening programme (in the period 1 April 1981 - 31 March 1991). The dataset used was constructed by combining data from:

- The central population register to identify women living in the municipality of Copenhagen at any time in the relevant time period and aged 50-69 years as well as the date of death or emigration
- The mammography register to obtain information about when women were first invited to screening
- The cancer register to obtain information about the date of first breast cancer diagnosis (women with breast cancer prior to first date of invitation were excluded)
- The cause of death register to obtain information about breast cancer as the cause of death

To analyse the effect of screening they compared the breast cancer mortality rates in the study groups with rates in the control groups, adjusting for age, period, and region. Analyses were performed using a Poisson regression model in which the breast cancer mortality rate can be written as a product of main and interaction effects:

$$\begin{aligned}\lambda_{epr} &= \theta \times (\theta_{exp})^{e} \times (\theta_{per})^{p} \times (\theta_{reg})^{r} \times (\theta_{exp,per})^{ep} \\ &\quad \times (\theta_{per,reg})^{pr} \times (\theta_{exp,reg})^{er} \times (\theta_{exp,per,reg})^{epr}\end{aligned}$$

where θ are main- and interaction effects of exposure (exp, 1=invitation to screening, 0=no invitation), period (per, 1=prior to screening, 0=during screening) and region (reg, 1=regions outside screening, 0=Copenhagen). The estimate of interest is the effect of exposure in the screening region in the screening period compared with no exposure in the screening region in the screening period, i.e.

$$\frac{\lambda_{e=1,p=0,r=0}}{\lambda_{e=0,p=0,r=0}} = \theta_{exp}$$

However, an estimate of $\lambda_{e=0,p=0,r=0}$ is not directly available. What can be estimated directly is the breast cancer mortality rate of the study group ($\lambda_{e=1,p=0,r=0}$), the breast cancer mortality rate of the national control group ($\lambda_{e=0,p=0,r=1}$), the breast cancer mortality rate of the historical control group ($\lambda_{e=0,p=1,r=0}$) and the breast cancer mortality rate of the historical national control group ($\lambda_{e=0,p=1,r=1}$). These give the following relative risks:
Study group rate/national control group rate=

$$\frac{\lambda_{e=1,p=0,r=0}}{\lambda_{e=0,p=0,r=1}} = \frac{\theta_{exp}}{\theta_{reg}}$$

Historical control group rate/historical national control group rate =

$$\frac{\lambda_{e=0,p=1,r=0}}{\lambda_{e=0,p=1,r=1}} = \frac{1}{\theta_{reg} \times \theta_{per,reg}}$$

and, therefore, the ratio of the two relative risks is

$$\theta_{exp} \times \theta_{per,reg}$$

It is not possible to estimate these two terms independently, i.e. the estimated exposure effect in this model is confounded by the interaction term between region and period. When using this approach for estimating the effect of the combination of invitation to screening and the interaction term between period and region adjusted for age, period, and region, the relative risk was 0.75 [95% CI: 0.63 ; 0.89]. For the county of Fyn the estimated relative risk was 0.78 [95% CI: 0.68 ; 0.89] (Njor et al. 2015).

3 The Weedon-Fekjær et al. (2014) approach

In Norway the national mammography screening programme was gradually implemented during the period 1995-2005. As in Denmark biennial invitations were sent to women aged 50-69 years. Weedon-Fekjær et al. (2014) also applied a Poisson regression model to evaluate the effect of invitation to screening. Women with a diagnosis of breast cancer after the first invitation date to mammography screening were regarded as exposed to screening, while women with a diagnosis of breast cancer before the first invitation date were regarded as unexposed. The Poisson model was used to compare incidence based breast cancer mortality among women invited to screening with those not invited while adjusting for age, effects of birth cohort, calendar time and county of residence. Furthermore, the estimated proportion of

breast cancer deaths that was attributed to breast cancers diagnosed after first screening invitation was added to the model as an offset. In the estimation of this proportion they used the interval from diagnosis until death from breast cancer among women who were not yet invited for screening. When applying this approach the estimated relative risk was 0.72 [95% CI: 0.64 ; 0.79].

References

Weedon-Fekjær, H., Romundstad, P.R. and Vatten, L.J. (2014). Modern mammography screening and breast cancer mortality: population study. *British Medical Journal*, **348**, 10.1136/bmj.g3701

Olsen, A.H., Njor, S.H., Vejborg, I., Schwartz, W., Dalgaard, P., Jensen, M., Tange, U.B., Blichert-Toft, M., Rank, F., Mouridsen, H. and Lynge, E. (2005). Breast cancer mortality in Copenhagen after introduction of mammography screening: cohort study. *British Medical Journal*, **330**, 10.1136/bmj.38313.639236.82

Njor, S.H., Schwartz, W., Blichert-Toft, M. and Lynge, E. (2015). Decline in breast cancer mortality: How much is attritutable to screening?. *Journal of Medical Screening*, **22(1)**, pp. 20–27

Do you have enough data? Things to learn from learning curves

Martin Sørensen and Kaare Brandt Petersen
SAS Institute, Denmark

Symposium i Anvendt Statistik, January 2016

Abstract

In this paper we construct learning curves for six different supervised learning models in order to answer the question *Do we have enough data?* and to reveal different fundamental properties of these six models. The results are based on a single data set called *Organics* and the six models are logistic regression, decision tree, neural network, k-nearest neighbours, support vector machine and gradient boosting.

1 Introduction

In classical hypothesis testing problems, one can in advance of the data collection answer the question *How much data do we need?* This enables an experimental design in which medical researches can set up tests of sufficient scientific depth and survey companies can set up cost efficient opinion polls. But in classical machine learning this is much harder. Before we have the data it is almost impossible to know if we need 500 or 5000 or 50000 observations to estimate a predictive model with reasonable precision. In this paper, we address the second best approach: When we have the first 500 observations and we are asked if we need more, then we can give an answer using so-called learning curves – we can answer the classical question *Do we need more data?*

2 Overfitting and data partitioning

In *supervised learning*, the aim is to estimate a function $y = f(\mathbf{x})$ which in the best possible way predicts the output y, given the multidimensional input $\mathbf{x}$. Sometimes y is referred to as a *target variable*, a *response variable* or a *dependent variable*, while $\mathbf{x}$ may be referred to as *input vector*, *treatment variable* or *independent variables*. To estimate the function, there is a set of example-pairs of input and output, given by the data set D

$$D = \{(\mathbf{x}_i, y_i) | i = 1, ...N\}$$

The function is said to *learn* the relation between input and output using the data set of pairs, which can be seen as given by some teacher or supervisor – hence the term *supervised learning*.

Since we rarely know what kind of function family is suitable for the task, we have to work with very flexible (large) function families, which are able to adapt to almost any pattern found in data. The flexibility, however, has a downside too: It is not only possible for the function

family to adapt to the essential dynamics of the problem. It is also able to adapt to every single uninteresting detail of the specific measured data set. Doing that, the model will not predict well in the future. The effect is called *overfitting* and is well known in machine learning [2].

To counter overfitting one has to both find the right function parameters and to adjust the function flexibility. If we denote the traditional function parameters by θ and a handle for control of function flexibility by λ, we have the function

$$y = f(\mathbf{x}; \theta, \lambda)$$

Everything we know about the relation between $\mathbf{x}$ and y is given by the data set D, so we have to use that and the simplest way is by so-called data partitioning: In data partitioning one randomly splits the data set D into three smaller (non-overlapping) data sets: training data D_{trn}, validation data D_{val} and test data D_{tst}. In a typical usage, a wide range $\lambda_1, \lambda_2, ..., \lambda_L$ of possible flexibility values λ is tried, and for each of these steps, the optimal θ_D is found using the training data D_{trn}. After that, the performance of the L models are measured using the validation data D_{val} to find the optimal flexibility λ_D. The performance of the function $y = f(\mathbf{x}; \theta_D, \lambda_D)$ is finally measured on the test data set D_{tst}.

This approach is standard in machine learning [2, 3]. The split between training, validation and test is often set to 40% for training, 30% for validation and 30% for testing, but many different variants are used, along with more advanced approaches such as cross-validation, leave-one-out or bagging [3]. The specifics of how to control the flexibility of a given function family is a topic in itself and can be chosen in many different ways. Very often, however, it has a component of input variable selection: A model with only a few input variables will almost always be simpler. In the literature, the flexibility is often referred to as model *complexity*.

In summary, for any function family f and data set D, there is a well defined way to find the best possible predictive function $y = f(\mathbf{x}; \theta_D, \lambda_D)$ and *best* here is defined as the lowest prediction error on future data sets.

3 Learning Curves

In machine learning, a *learning curve* is the plot of performance (learning) as experience increases. Experience can mean several things such as number of observations in the data set or number of iterations in the learning algorithm.

In this paper, learning curve means a plot of performance (ROC index) as fractions from 1% to 100% of the available data set of size N is chosen and used to find the best possible model. Although, from a formal viewpoint, only the test data can be used to express the model performance (learning), we are in this paper computing learning curves for training performance, validation performance and test performance. The reason is, as we shall see later on, that interesting properties of the model can be seen by comparing the performance measures of training, validation and test data set.

4 Example data set

The data set used is a data set called *Organics* from the SAS education material. It contains 22223 customer interactions in a supermarket chain and can be used to train a model to predict if a customer is likely to buy organic products or not. To do that, there are $N = 22223$ historical customer interactions with a variable expressing if the customer bought organic products or not along with knowledge about customer age, gender, geographical region, television region, and

Variable	Type	Role	Information
Affluence grade	Interval	Input	Min=0, Max=34, Mean=8.71
Age	Interval	Input	Min=18, Max=79, Mean=53.80
Total spend	Interval	Input	Min=0.01, Max=296313.85, Mean=4420.59
Loyalty Card Tenure	Interval	Input	Min=0, Max=39, Mean=6.56
Neighbourhood	Class	Input	Levels=7 (A,B,C,D,E,F,U)
Gender	Class	Input	Levels=3 (M,F,U)
Geographic region	Class	Input	Levels=5 (Midlands, North, ..)
Television region	Class	Input	Levels=13 (Border, London, East, ..)
Loyalty status	Class	Input	Levels=4 (Gold, Platinium, ..)
Organic Purchase	Class	Target	Levels=2 (0 or 1)

Table 1: The Organics data set. The data set is 22223 instances of super market customers either buying organic products (Target=1) or not buying organics (Target=0). In order to predict if the customer will buy or not there are 9 input variables.

others – se Table 1 for a list of variables. For the purpose of the learning curves it suffices to say that it is a binary target problem with unbalanced classes: Target=0 in 16718 observations (75.23%) and Target=1 in 5505 observations (24.77%).

5 Model Gallery

The models (function families) used in this paper are very diverse and it goes far beyond the scope of the paper to describe them in details. In stead, they are introduced on a conceptual level with a few lines per model and with references to where one can read more about each of the models. All models are implemented in SAS Enterprise Miner.

Regression: A traditional logistic regression model is used with the standard cost function and no variable selection. Only main effects are included. Read more about logistic regression models in [5].

Decision Tree: Data driven decision trees work by successively dividing the data into smaller groups of similar target values. For every possible split of the data, the algorithm includes all variables and test their relation to the target variable. Each split is then at the variable and value which has the strongest relation to the target variable. Read more on decision trees in [6].

K-Nearest Neighbours The k-nearest neighbours model is a model in which, for every data point to predict, the nearest k neighbours in the training data set is found, and used to find out the most likely prediction – typically by voting (if the majority of neighbours belong to class 1, then class 1 is chosen and vice versa). In this paper $k = 16$ is used. Read more in [2].

Neural Network: Neural networks are a type of algorithms using a mix of linear and non-linear transformations to produce the output. The construction allows for a great flexibility which enables the network to adjust to many different patterns. In the specific case of this paper, a two-layer network perceptron is used and in this hidden layer there are 3 hidden units as well as bias units. [1]

Support Vector Machine Support Vector Machines (SVM) attempt to find the examples (observations) in data which distinguishes the classes from each other. It does so by mapping the problem of non-linear two class problem in the low dimensional input space (in this case 9D) into a linear problem in an infinite dimensional space. Al though this sound unpractically complicated, it turns out that using the so-called *kernel-trick* one does not need the explicit mapping but rather the inner products of the mapping which can be approximated by gaussian densities. Read more about SVM's in [2].

Gradient Boosting: Boosting is a technique in which a sequence of models (refinements) are estimated such that the subsequent models are correcting, as best possible, the predecessors misclassifications. The data points at which predecessors are mistaken are *boosted*, i.e. given higher weight to ensure improved overall importance. In this paper is used an implementation in SAS Enterprise Miner, in which the models in the sequence are decision trees. Read more in [3].

6 Results

For all the models presented in the previous section, the optimal model has been estimated for various sizes of data set. The original data set consists of $N = 22223$ observations, but in order to construct the learning curves this data set is subsampled randomly. The smallest data set $D_{1\%}$ is 1 % of the original data, i.e. $N_{1\%} = 222$, the second smallest $D_{2\%}$ has the size $N_{2\%} = 444$, and so forth up to $D_{100\%} = D$. By this we get three learning curves – one for performance on each of the three data parts training, validation and test – at 1%, 2%, ..., 99%, 100% of the original data set.

As performance measure on this binary class problem is chosen Receiver Operator Curve (ROC) index. The ROC index is a performance measure and not an error measure (that is, *higher is better*), and is always between 0.5 (random) and 1.0 (perfect ranking).

Learning curves for one model: Figure 1 shows the learning curves for the logistic regression model. Note especially:

- Following the curve of the test data set it is clearly influenced of some sort of random noise, making it fluctuate. This is due to the random data selection and partitioning for each subsampling. Sometimes specific data points are included or excluded which influences the performance.

- The performance on the validation data set and the test data set are indistinguishable over the entire range of data set sizes; thus in this case there is no overfitting by the validation set. The training performance, however, is better than the two others in the range from 1% to approximately 10% of the data set size. This is a signature of risk of overfitting: If we had not been doing the data partitioning and instead used a single data set of this size, we would have misjudged the performance and ended up with a suboptimal model.

- From about 30% and to 100%, the test performance is not increasing, that is, for this specific data set and this specific model, there is a data saturation at approximately $N = 6600$ observations. Thus, in this case, *yes* we have enough data.

Comparing six models. Figure 2 shows the learning curves for all six models. Many interesting aspects can be read from this figure:

- The neural network has much the same overall signature as the logistic regression but much higher training performance, meaning that the risk of overfitting for neural networks are higher. This makes sense since the neural network includes the logistic regression as a special case but is much more flexible as function family.
- The support vector machine (denoted HPSVM) has an overall signature similar to the logistic regression, only less pronounced. That is, support vector machines can overfit too, but in this data example, it is not as sensible to data set size, until it gets very small. This may be because of methods approach in which so-called support vectors along the decision boundary between target=0 and target=1 are identified; as long as there are enough support vectors to define the boundary, no real change is happening. Note also, that the SVM seem to be the most robust of the six: The fluctuations are smaller and the measures of the three different data partitions are coinciding for most of the range.
- The k-nearest neighbour model (denoted MBR) has an entirely different structure: The training set performance is consistently far ahead of the others. This is due to the nature of the nearest neighbours algorithm: When the data point itself is a part of the data set for deciding the class (as is the case for the training data set), it will always be in the center of the estimation, because a data point always is a close neighbour to itself. Another special property of the MBR learning curves is that they are increasing over the entire range: in this case we don't have enough data – or at least we would get better results if we had more data.
- The decision tree as a special signature as well: Firstly the performance of all three data partitions follow each other closely, thus giving no signs of risk of overfitting, and secondly, the training performance is not increasing for small data sets, but rather decreasing. The reason is that for small data sets (at least for this specific data set), the decision tree is becoming a rather simple model with very few end-nodes and a decision tree like that having only, say, 4 end-nodes will have decreased performance on the training set also.
- The gradient boosting model has a signature similar to the logistic regression, but that is actually at first sight surprising because it is a sequence of decision trees. The simplicity of the decision trees of small data sets does not hold for gradient boosting models, however, since there will be a number of end-nodes for each tree in the sequence, thus in total producing a higher number of end-nodes even for small data sets. Note also the fluctuations are much smaller.

One property all models except the k-nearest neighbours have in common is that above 50% of the data set, the test set performance is not improving and we can for this data set safely conclude that *yes, we have enough data.*

7 Summary

In summary, we have presented learning curves for six different models on the same data. The results demonstrate that learning curves are practical approaches which can be applied to any data set to answer the question *Do we have enough data?* Furthermore it is illustrated that learning curves also are very interesting for understanding fundamental aspects of different models for supervised learning.

References

[1] "Neural Networks for Pattern Recognition", Christopher Bishop, Oxford University Press, 1995.

[2] "Pattern Recognition and Machine Learning", Christopher Bishop, Springer, 2006.

[3] "Stochastic Gradient Boosting.", Jerome Friedman, Computations Statistics & Data Analysis 38: 367-378, 2002

[4] "The Elements of Statistical Learning. Data Mining, Inference and Prediction", Trevor Hastie, Robert Tibshirani and Jerome Friedman, 2. edition, Springer 2009.

[5] "Applied Logistic Regression", David Hosmer, Stanley Lemeshow and Rodney Sturdivant, 3. edition, Wiley 2013

[6] "Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner", Barry de Ville, SAS publications 2006.

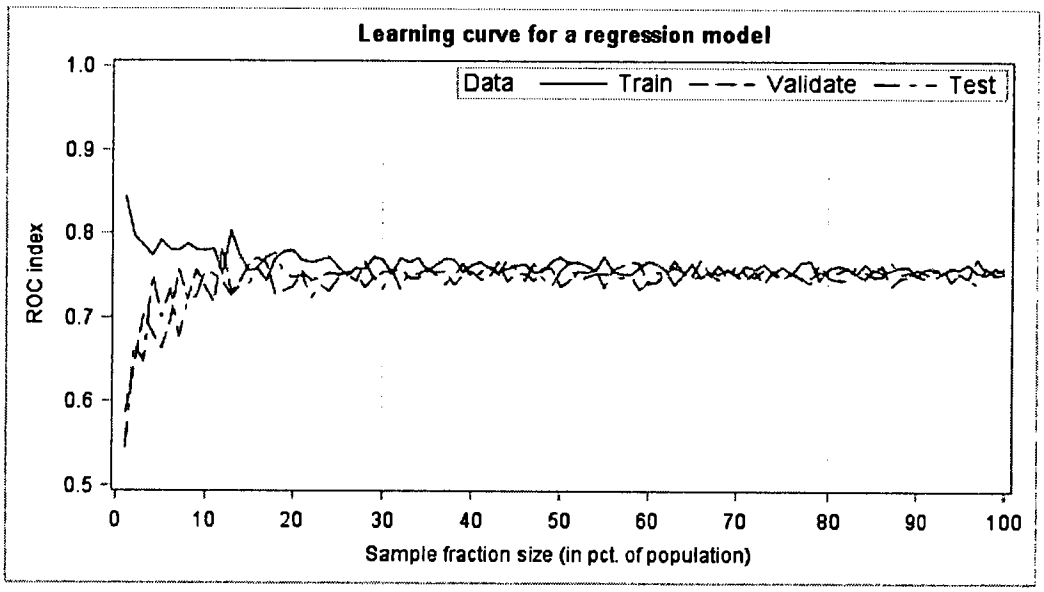


Figure 1: Learning curves for the logistic regression model. Note that the training performance is higher for smaller data sets, which is an indication of potential overfitting. Note also, that the test performance is not increasing over a long range, thus indicating, that in this case we have enough data.

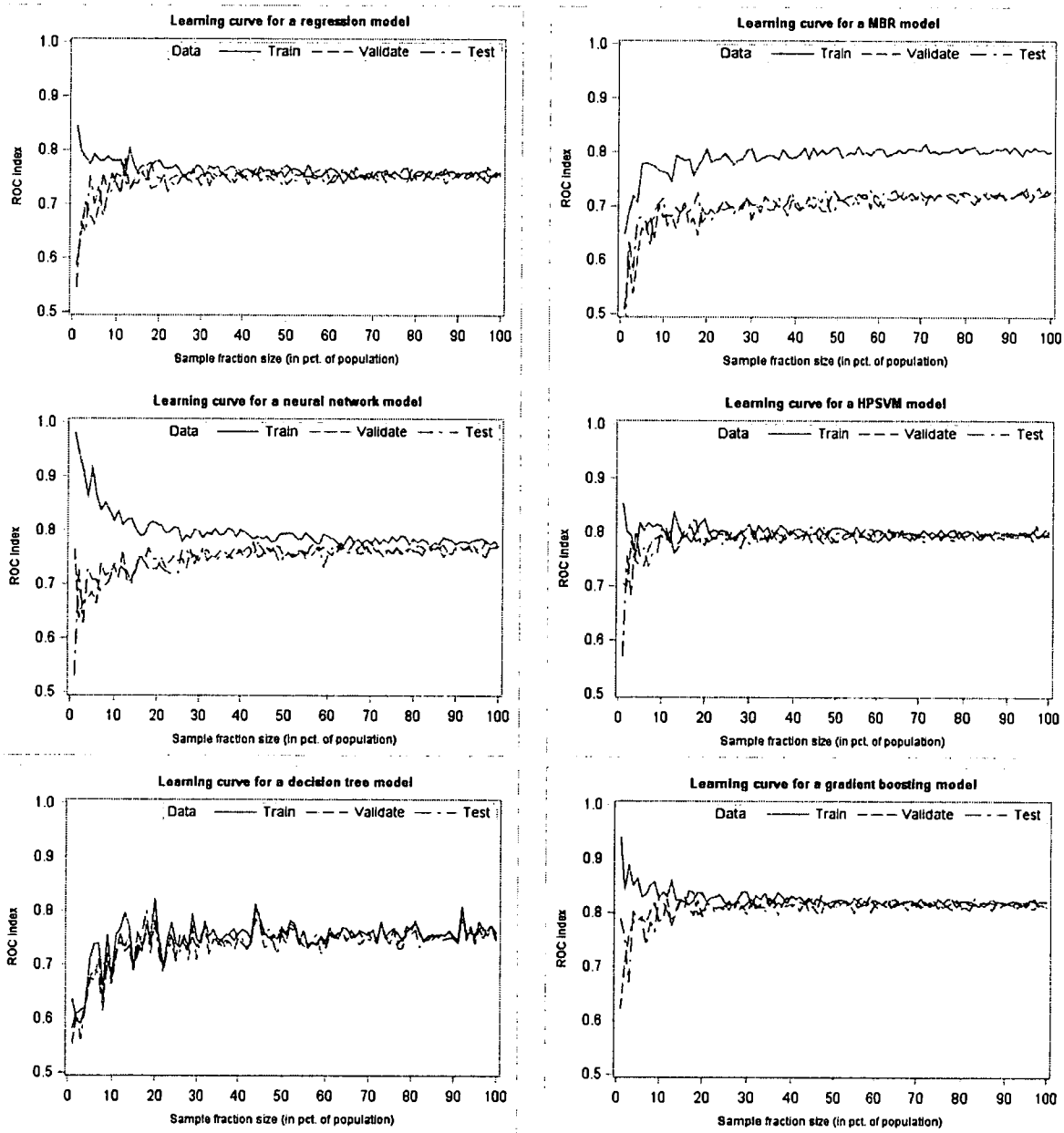


Figure 2: Learning curves for six different models: Logistic regression, K-nearest neighbours, Neural network, Support Vector Machine, Decision tree and Gradient Boosting. See the text for comments on the results.

THE POWER DISTRIBUTION AS A MODEL FOR CRIMINAL CAREERS

Thomas Lill Madsen

Abstract The length of criminal careers occur at a frequencies similar to drawings from a power distribution. Knowledge about the s-parameter of the distribution is considerably more informative than its mean value, known to criminologist as 'lambda'. Crime rates are truly decomposable as participation times frequency, with participation determined by demography and lambda generated by a stochastic power process.
Keywords power laws; clearance rate; charges per person; participation; lambda; demographic crime model

---

1. A criminal career is the sequence of offences during some part of an individuals lifetime. For the individual involvement in crime has two aspects: to participate or not, and if yes, with which frequency. Criminologists have had good success with models for participation but less so with frequency. Here criminologists use lambda, the mean number of crimes committed (or arrests made) per active offender per year, as one of the primary measures of criminal career intensity. Lambda recommends itself mainly by ease of calculation, but in spite of numerous texts it remains unclear of which *distribution* lambda is the mean- literally what lambda *means*. Thus for a given value of lambda, say 2, criminologists have no clue as to what this value predicts for the number of careers with 1,2,..,n crimes.- The objective of this paper is to throw light on precisely this subject by for the first time introducing power laws as the essential model for criminal careers.

A large number of natural and social phenomena display analogous laws. In particular power laws have been found applicable to biology, geography, physics, economics (Bak 1997). Introduced by the Italian economist V. Pareto (1848-1923) in 1906 as way of describing the upper reaches of the income distribution, the power distribution has become recognized thanks to its ubiquity.

The power distribution is in particular characteristic of systems in a critical state, on the border between order and chaos. In this state, *cascading* is likely to occur, i.e. one event paving the way for the next. Such distributions are characterized by many small events and few large ones. Examples are the distribution of cities with X inhabitants, number of occurrence of specific words in a text, number of sex partners through life and the tectonics of the Earth's crust generating quakes of magnitude X.- The bright guys from the physics department were the first see the universality in this (Bak, 1997).

2. Previously it has been theorized that the once a career is initiated by its first act, further crime generation is shaped by a poisson process with lambda as its mean (Blumstein, Petras). This implies a static, memoryless, process with constant probability of committing a further crime the next day. In contrast, the power distributed criminal is dynamic and state dependent in the sense that after commiting N crimes, the probability of moving forward to N+1 is an increasing function of N, as shown below. Thus if one crime increases the likelihood of doing yet another, the power formula provides a framework for capturing the dynamics and unfolding of careers.

We can abstractly represent a criminal career by the number crimes, N, committed within a given time horizon. For a power distribution $p(N) = N^{-s}$,where p(N) is the frequency of N= *1,...,inf,* and *s* is the parameter of the distribution, typically in the interval [2,3]. On a log-log scale p(N) plots as a decreasing linear function of N with slope *-s.* It gets more and more likely to move on from N to N+1, as the ratio p(N+1)/p(N) is an increasing function of N, i.e.: $P(N+1|N) = N^s / (N+1)^s$. Eg. for N = 1 and s= 2: $p(N+1) = 2^{-2} = .25$ whereas for N = 10, p(N+1|N) = 100/121 = .8- A slippery slope indeed *under the assumption of unrestrained movement* from N to N+1. In practice, the law will interfere at some stage, apply sanctions and thereby dampen the process at the extreme.

3. The demographic descriptors, age/sex/ethnic origin, are highly significant factors for modeling *the crime rate* of a population segment. This measure is by definition decomposable as:

crime rate = participation rate * offender frequency

A central question has been whether demography contribute to career length for those participating, ie. charged at least once. If the participation rate is determined by demography,

either the same factor contribute to frequency as maintained by Petras(2009) or, hypothetically, power laws take over and generate frequency by a *stochastic* process.

Data and method

Two set of data from Danish police records were available for this study. Set *A* containing data for all 106000 year-persons charged for 185000 crimes over the two years 2005-6. All in all, 71380 perpetrator were indicted. Set *B* is semi-longitudinal and contains similar records for the cohort of '85 (N = 58000) from their legal debut at age 15 in 2000 until mid 2007. For each person in the two sets, the records contain the count of charges per year and crime type over the period. Dataset B will provide provisory indication as to the difference between the outcome of a single year vs. a somewhat longer period of time.

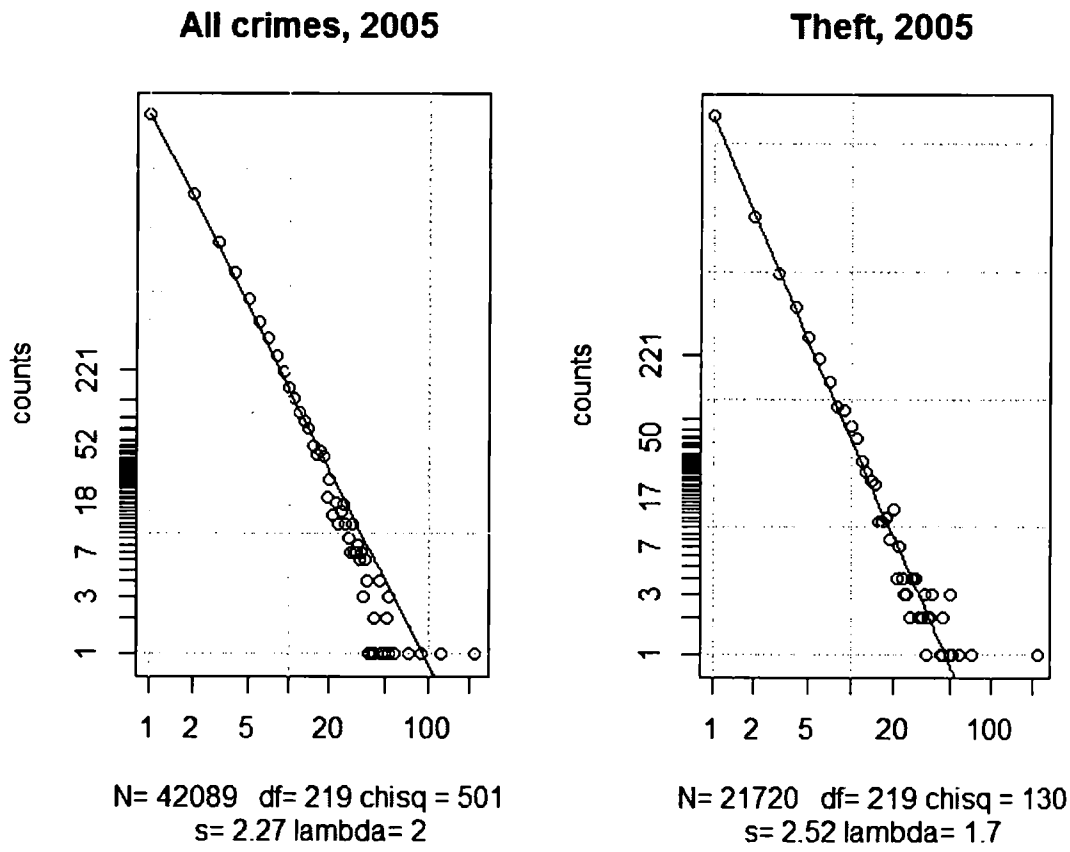
(1) The distribution of the variable 'career length'

The data is set up for analysis by counting the number of persons charged 1,2,...,N times. If criminals do not to specialize, (Gottfredson 1990), a criminal career over a given time period is simply the count of all the crimes committed for a person, whatever the crime type. Alternatively, sub careers can be analyzed by further sorting down to person/crime type/year as the basic unit.

With these data it is now straightforward to a model a power process with the parameter *s* to be estimated by regressing a) the numbers of careers of length 'i' on b) 'i', in a log-log poisson regression. The two coefficients that emerge are: a constant term reflecting the number of careers, and *s*, the slope parameter.- Farrington (1992) alluded to the potential relevance of *a crimes serial number* for predicting the time interval to next offence. Here it is used in a similar vein.

To illustrate, the data in fig.1A shows for data from A-set. It displays the count of persons on the y-axis and the number of crimes on the x-axis. We see many brief careers and few large ones, the later occurring predominantly only once (counts = 1). The predominant case is *one* charge per person. For N< 20 the fit seems quite good but with a tendency to over predict for longer careers. To assess the model fit more formally we can use the chisq-statistics vs. degrees of freedom. Ideally chisq should be in the range of df +/- $(2*df)^{.5}$. Here we have chisq = 501 and df = 219, ie. a not so good fit. For sub careers in theft the fit is excellent.- A separate question is whether the timeframe makes a difference.

Fig 1. Power laws at two levels, 2005.



(2) The two stage process

A central question is whether individual factors, eg. demography, add information in career terms for those who *already* are charged at least once, ie. participating. Blumstein (1987) and Farrington (1992) concluded negatively on this whereas Hirschi and Gottfredsson (1990) and Petras (2009) claimed that criminogenic factors have an influence on both participation and frequency.

In order to model both crime rate and participation, we can try the basic demographic crime model:

(2) log(*crime*) = age + sex + ethnic origin

where 'sex' and 'ethnic origin' are level constants (binary). The formula presumes that crime declines exponentially with age when grouped by 5-year intervals. There are two versions for (2):

1) *crime* equaling *crime rate* and

2) *crime* equaling *participation*

To the extent that the parameters from 1) and 2) are similar, crime rate is proportional to participation and it will be a vindication of the two stage process: demographic parameters determine *participation*, and *frequency* takes care of itself by a stochastic process.

Equation (2) assumes that crime rates declines exponentially by age, which turns out to be a fair approximation when using 5-yr intervals, ie. crime(x) = k0*exp(*a*x), where x is age and k0 and *a*, the decline rate, are the parameters to be estimated. Parameter k0, the level constant, is allowed to differ by ethnic origin and sex (*a* also varies by sex but this detail is left out for simplicity).

Results

A. Power laws for major crime types

Expanding on fig.1, **fig.2** and **table 1** summarizes empirical results for sub career length frequencies.

Figure 2 Power laws by crime type for two year period 2005-6.

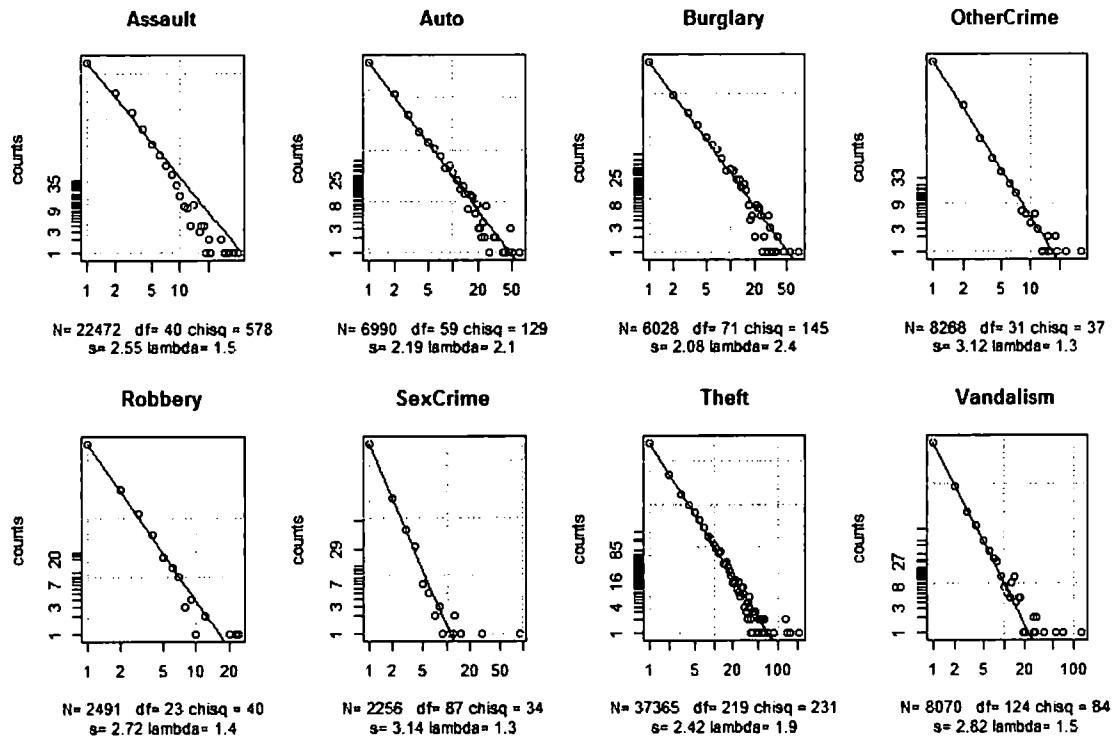


Table 1. Parameter summary for criminal careers over two years

	lambda	s	df	chisq	N
Assault	1.5	2.5	40	578	22472
Auto	2.1	2.2	59	129	6990
Burglary	2.4	2.1	71	145	6028
OtherCrime	1.3	3.1	31	37	8268
Robbery	1.4	2.7	23	40	2491
SexCrime	1.3	3.1	87	34	2256
Theft	1.9	2.4	219	231	37365
Vandalism	1.5	2.8	124	84	8070

For a goodness-of-fit test in this table we compare df, degrees of freedom, with chisq. In general, the fit is not bad. The s-parameter ranges from 2.1 to 3.1. The largest residuals appear to be in *assault* (negative for larger x). In this crime apprehension risk is high (visual contact with the perpetrator and police priorities) and can lead to immediate incapacitation and an artificial stop to the unfolding of the career. Same tendency but in smaller measure is seen for Auto and Burglary.

The lower the value of *s,* the higher the prevalence of the industrious perpetrator. Burglary and Auto crime have the lowest values, reflecting the reiterative nature of criminal behavior in these areas. But both the specific and relative values of s should however be treated with some circumspection due to the skewing influence of different clearance rates ('sandsynlighed for opklaring'): everything else being equal, a lower clearance rate will lead to a higher value of s. To intuitively see why, consider that as the clearance rate approaches zero, all known perpetrators will be preeminently be facing only one charge, ie. leading to a high value of s. In other words, the s-values shown in table 1 would be lower if all crimes were solved but by differential amounts depending on clearance rates.

The s-parameter depends further on the time frame and for an analysis of this relation a longitudinal dataset is required. Data for cohort '85 from onset in 2000 to mid 2007 allows comparison of 1-yr vs. 7.5 years time frames (the best available at the time). The data refers to 8546 persons, some charged in one year only, some in several. In fig. 3A-C. Fig. 3A shows results for 20700 1-yr sub careers and fig. B shows the same data but now for the whole period for the same 8546 persons.

In fig. 3A it is again evident that power distribution is a good model for yearly data; also note in comparison with fig.1 that the estimates of the s-parameter and lambda are similar. In fig. 3B, lambda has increased from 1.7 in to 4.1 and *s* correspondingly decreased to 1.79, implying increased dominance of the habitual criminal. For the 7-year period the model still does well for N < 30 but over predicts the number of very long careers. In this connection Farrington (1992)

introduced the hypothesis that *desistance* occurs continually at a certain rate. This is similar to complementing the model with a term for exponential decay whereby it then reads: $\log(p(N)) = \log(N^{-s}) + N$. Fig 3C show the implementation of this idea which visually and by chisq-statistic seems to restore the model fit.- Whether this term reflects true desistance, less street time or other restraining influences from the juridical system is untold.

Fig 3. Power laws for crime over different time horizons.

Fig 3A. Sum of all crimes 2004, yearly

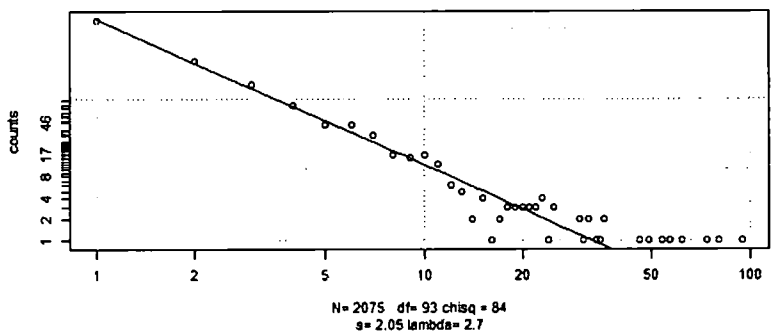


Fig 3B. Sum of all crimes 2000-7, in toto

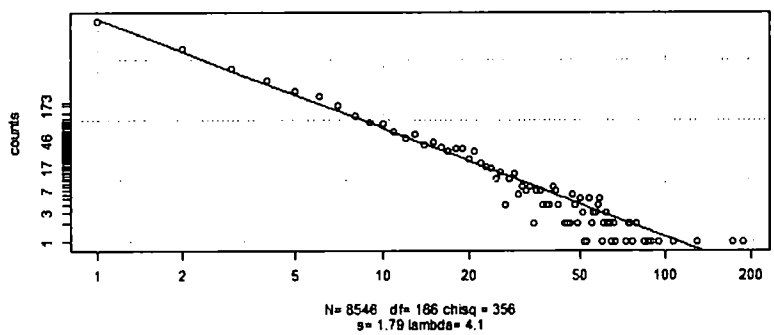
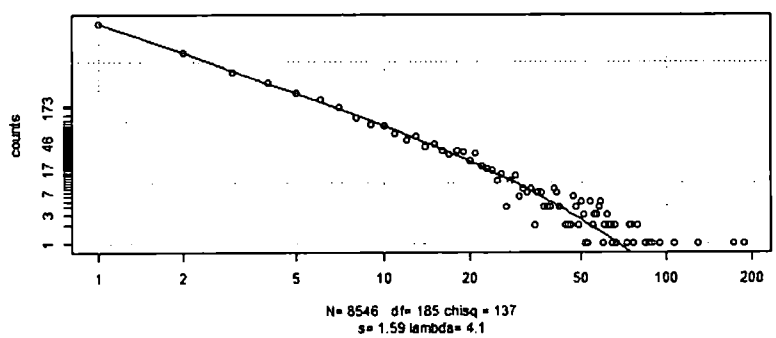


Fig 3C. 2000-7 combined and with exponential term



B. Decomposition of crime rate, Blumsteins model

A modeling of the age-crime curve in a demographic model with respectively crime rate and participation as the crime measure, yields the following estimates for the period 2005-6:

Table 1. Age-crime-curve parameters:

A.Estimation for *crime rate*:

	Estimate	Std. Error	z value
(Intercept)	-1.886	0.009	-207.483
x	-0.066	0.000	-343.746
sexm	1.721	0.007	264.436
origin_west	-0.930	0.006	-160.311

B. Estimation for *participation*:

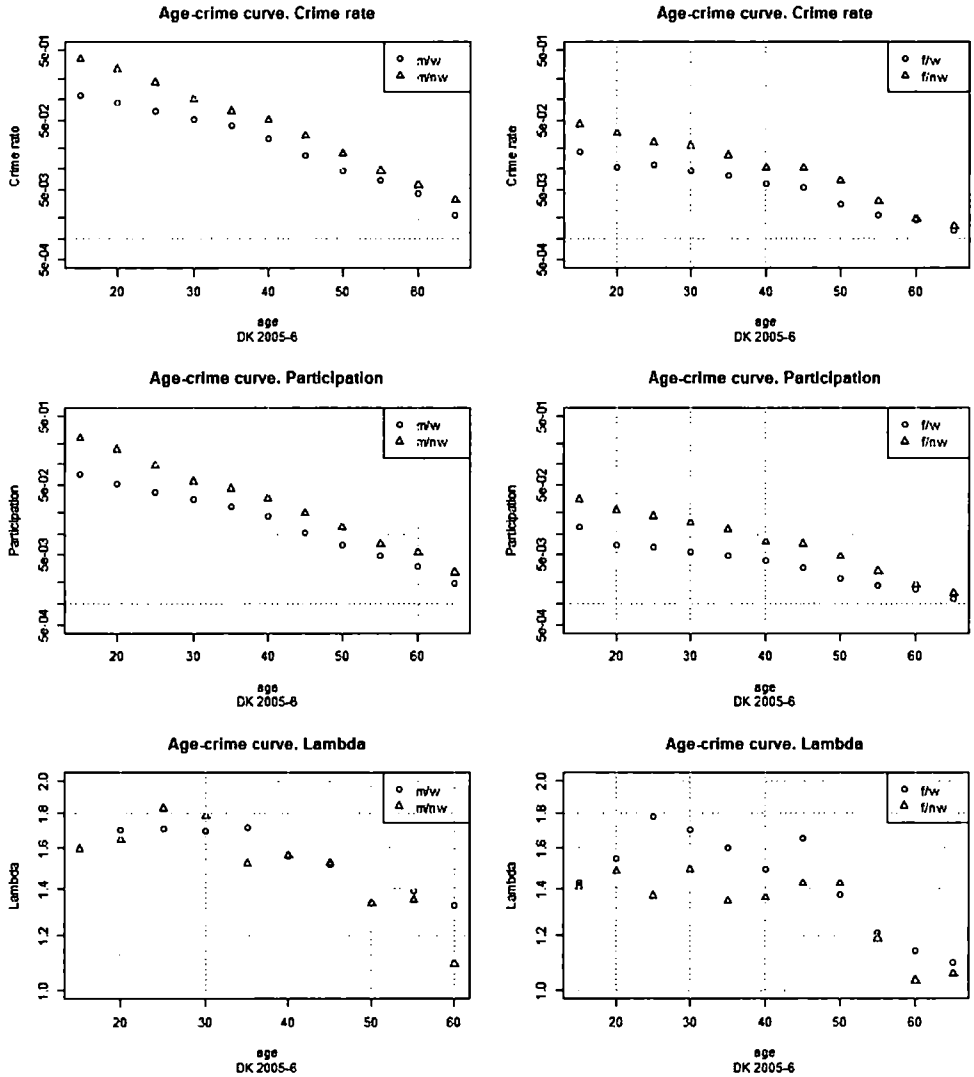
	Estimate	Std. Error	z value
(Intercept)	-2.365	0.011	-208.714
x	-0.063	0.000	-263.774
sexm	1.637	0.008	204.213
origin_west	-0.936	0.007	-127.441

Reading from table A, the parameter for 'x' indicates the decline rate per year of age. Ie. crime rate declines 6.6% per year. For western origin crime rate is a factor exp(-.93) ~ 2.5 lower, and for males vs females a factor exp(1.72) ~ 6 higher. Fig. 4 shows ever declining crime rates from the onset when grouped by 5-year age intervals.

Comparing the coefficients of A vs. B in table 1, the intercept is sui generis different as it is an estimate of lambda as exp(z) where z is respectively 2.365- 1.889. ie. lambda = 1.61.

Overall, the interesting parameters are virtually identical when comparing A to B, which is also seen by lambda being rather constant over the more interesting age groups (15-45). The implication must be that *demographic variables only influence the probability of participation* and that those who participate are alike in behavior. Comparing ethnic Danes vs. immigrants, the later have a crime factor of at least 2.5, but the *perpetrators* from both group have identical behavior in terms of lambda and choice of crime type except for robbery.

Fig. 4. Age-crime curves by origin (w/nw) and sex (m/f).



Discussion

The analysis has indicated that the power distribution works best at the sub career level but for the entire career has a tendency to overshoot for $N > 20$. It appears amendable to a fix by the exponential extension, but it remains to open if this has any ontological substance, eg. is a result of incapacitation, or is just to be seen as an adhoc measure.

The analysis has been based on crimes with known perpetrators but these constitute a minor fraction, about 1/5, of all crimes. The power distribution can by simulation enable inferences about the *totality of crimes* committed and their distribution over known and unknown offenders. This could be a topic for further research.

Conclusion

The power distribution was one of the missing links in criminologists' arsenal and the essential model for frequency generation in criminal careers.. All though there is a monotonic relation between *s* and lambda, only the former represents operative knowledge in the sense that the distribution of career lengths can be predicted from it. From the count of subjects charged once and twice we can predict the number that will be charged *N* times even though these individuals and deeds are entirely unconnected. The stochastic nature of human conduct is underscored: unpredictable at the level of the individual but predictable *en masse*. - There's the power of statistics.

References

Bak, Per: *How Nature Works- the Science of Self-organized Criticality,* Oxford University, 1997

Blumstein, A., Cohen, J.: *Characterizing criminal careers,* Science, vol. 237,1987

Farrington, D.P.: *Criminal career research in the United Kingdom,* Brit.J of Criminology, vol.32/4, 1992

Gottfredson, M. & Hirschi T. *A general theory of crime,* Standford Univ. Press, 1990

Petras, H., Nieuwbeerta, P., Piquero,A.: *Participation and frequency during criminal careers over the life span.* 2009, www.researchgate.net/publication/45713492

Software: http://www.R-project.org.

Är SD Sveriges största parti?

Jakob Bergman & Björn Holmquist
Statistiska institutionen, Lunds universitet

Sammanfattning

Den 20 augusti 2015 hävdade dagstidningen Metro att Sverigedemokraterna var Sveriges största parti. Detta baserade man på att partiet blivit det största i en opinionsundersökning av företaget YouGov. Men hur kan man testa påståendet att en specifik andel är den största? Vi tar vår utgångspunkt i andelarnas speciella parameterrum simplex och dess inbyggda restriktioner. Vi visar hur man kan konstruera ett test för att avgöra om en specifik andel är störst baserat på en likelihoodkvotansats, där vi utnyttjar en isometrisk logkvotstransformation för att underlätta de numeriska beräkningarna. Eftersom man vid denna typ av problem typiskt enbart har en enda observation av de relativa partipreferenser, diskuterar vi teststorhetens fördelning. Vi illustrerar våra resonemang med data från den ovan nämnda undersökningen av YouGov.

1 Introduktion

Torsdagen den 20 augusti 2015 hade dagstidningen *Metro* rubriken "Nu är SD Sveriges största parti" över hela förstasidan (Wallroth, 2015). Från en journalistisk synpunkt var rubriken inte alls förvånande; för tio år sedan var Sverigedemokraterna (SD) ett parti som på sin höjd hade 1–2 % av väljarkåren och sällan, för att inte säga aldrig, ens rapporterades i opinionsundersökningarna, och nu förelåg en undersökning som gav SD den största andelen av väljarkåren av något parti. En förändring som saknar motstycke i modern svensk politik. Från en statistisk synpunkt var rubriken mera förvånande. Hur kunde *Metro* vara så säkra på att SD var det största partiet? Faktum var, att de tre största partierna enbart skilde ett par procentenheter åt. Så för att dra slutsatsen att SD var det största partiet fordrades någon form av hypotestest. Men hur testar man påståendet att en specifik andel är större än alla de övriga, givet en vektor av observerade frekvenser?

2 Parameterrummet för andelar i väljarkåren

Anta att det finns D partier ($j = 1, \ldots, D$) och varje väljare i väljarkåren tillhör ett parti. Vi låter vektorn $\boldsymbol{p} = [p_j]$ vara partiernas andelar av väljarkåren. Eftersom andelarna p_j är icke-negativa och måste summera till 1, så utgörs parameterrummet för $\boldsymbol{p}$ av simplexen $\mathbb{S}^D$. Om man tar ett obundet slumpmässigt urval om n väljare ur väljarkåren och låter $\boldsymbol{X}$ vara antal väljare för respektive parti, så kommer $\boldsymbol{X}$ vara multinomialfördelad med parameter $\boldsymbol{p}$. Baserat på vårt stickprov så önskar vi testa hypoteserna

$$H_0 : \boldsymbol{p} \in \mathbb{S}^D - \omega_i$$
$$H_1 : \boldsymbol{p} \in \omega_i$$

där ω_i är det underrum av $\mathbb{S}^D$ där parti i är störst, dvs. där p_i är den största andelen. Gränsen mellan de två underrummen utgörs av den linje, plan etc. där $p_i = p_j$ för något $j \neq i$ och p_i är större än alla övriga p_j.

Att parameterrummet är en simplex medför bl.a. att parametrarna är negativt korrelerade; om en andel ska kunna öka så måste minst en andel minska. Detta och de övriga restriktionerna i rummet medför att det ofta kan uppstå praktiska svårigheter vid beräkningar (se t.ex. Aitchison (1986) för fler detaljer). Aitchison (1982) introducerade logkvottransformationer som en lösning på en del av problemen. Det finns flera olika logkvotstransformationer, men den numera mest populära är den *isometriska logkvotstransformationen (ILR)* (Egozcue et al., 2003). Den innebär att problemet överförs från $\mathbb{S}^D$ till det reella rummet $\mathbb{R}^{D-1}$. Rent numeriskt finns det flera sätt definiera ILR. I denna artikel följer vi versionen i Egozcue et al. (2003) och låter $\mathrm{ILR}(\boldsymbol{p}) = \boldsymbol{y} = [y_j]$ där

$$y_j = \frac{1}{\sqrt{j(j+1)}} \log \frac{\prod_{k=1}^{j} p_k}{p_{j+1}^{j}}, \qquad j = 1, \ldots, D-1. \tag{1}$$

Om vi exempelvis bara har tre andelar $\boldsymbol{p} = (p_1, p_2, p_3)'$ så är följaktligen

$$\boldsymbol{y} = \left[\frac{1}{\sqrt{2}} \log \frac{p_1}{p_2} \quad \frac{1}{\sqrt{6}} \log \frac{p_1 p_2}{p_3^2} \right]'. \tag{2}$$

Exemplets parameterrum $\mathbb{S}^3$ med underrummet ω_1 illustreras som ett triangeldiagram i Figur 1a. I Figur 1b visas det motsvarande reella rummet $\mathbb{R}^2$ med motsvarande underrum $\omega_1^* = \mathrm{ILR}(\omega_1)$. X-axeln utgörs här av y_1 och y-axeln av y_2 i (2). För fler detaljer om ILR och dess egenskaper, se Pawlowsky-Glahn et al. (2015, kap. 4). Vi nöjer oss här med att konstatera att det finns en invers ILR-transformation sådan att $\boldsymbol{p} = \mathrm{ILR}^{-1}(\boldsymbol{y})$. Analytiska uttryck för den version av ILR som använts här är dock tämligen komplicerade. Som exempel följer den inversa transformationen av (2)

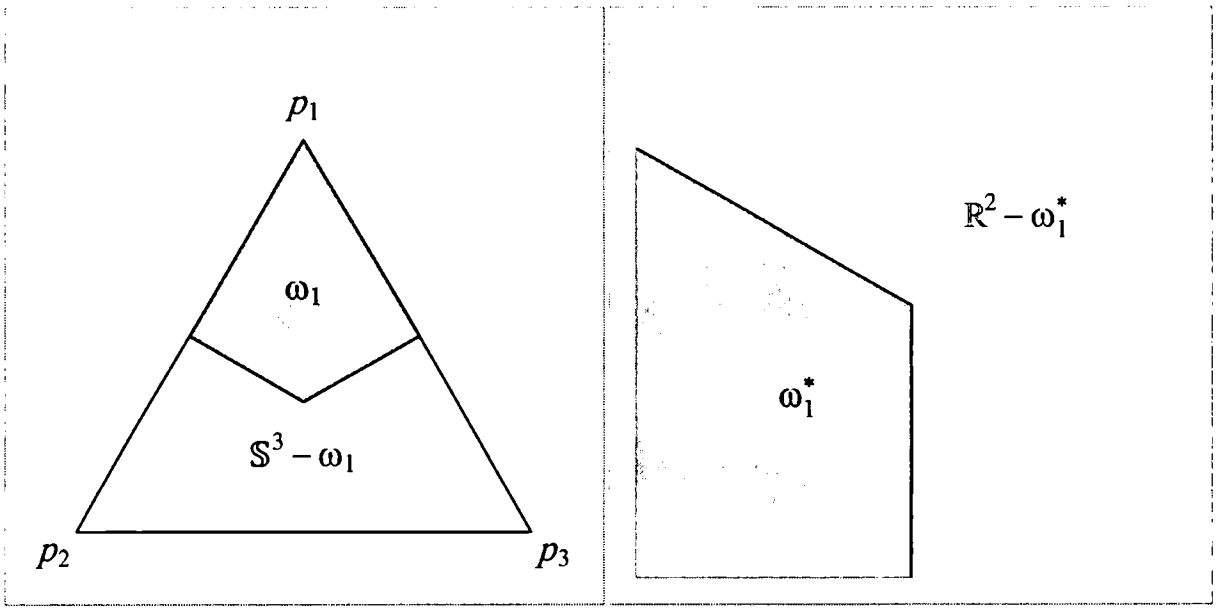
$$\boldsymbol{p} = \frac{1}{A} \left[\exp\left(\frac{y_1}{\sqrt{2}} + \frac{y_2}{\sqrt{6}} \right) \quad \exp\left(-\frac{y_1}{\sqrt{2}} + \frac{y_2}{\sqrt{6}} \right) \quad \exp\left(-\frac{y_2\sqrt{2}}{\sqrt{3}} \right) \right]'$$

där

$$A = \exp\left(\frac{y_1}{\sqrt{2}} + \frac{y_2}{\sqrt{6}} \right) + \exp\left(-\frac{y_1}{\sqrt{2}} + \frac{y_2}{\sqrt{6}} \right) + \exp\left(-\frac{y_2\sqrt{2}}{\sqrt{3}} \right).$$

3 Ett maximum likelihoodkvottest

I Sverige finns sedan 2010 åtta partier i Riksdagen. Utöver dessa finns det ytterligare ett eller två partier som har varit i närheten att komma över fyra procentsgränsen till Riksdagen. I Metro redovisas andelar för nio partier. Till dessa nio partier kommer slutligen alla övriga partier som normalt tillsammans samlar mindre än en procent av



Figur 1. I (a) illustreras parameterrummet $\mathbb{S}^3$ med underrummet ω_1, där p_1 är den största andelen, och underrummet $\mathbb{S}^3 - \omega_1$, där p_1 inte är den största andelen. Den övre spetsen av diagrammet utgörs av parametervärdet $\boldsymbol{p} = (1, 0, 0)'$, den nedre vänstra spetsen av $\boldsymbol{p} = (0, 1, 0)'$, och den nedre högra spetsen av $\boldsymbol{p} = (0, 0, 1)'$. Gränsen mellan de båda underrummen utgörs av linjen från $\boldsymbol{p} = (1/2, 1/2, 0)'$, via $\boldsymbol{p} = (1/3, 1/3, 1/3)'$, till $\boldsymbol{p} = (1/2, 0, 1/2)'$. I (b) visas motsvarande underrum i det reella rummet $\mathbb{R}^2$.

väljarkåren. Enligt Metro består således väljarkåren av tio olika andelar. Vi vill testa om SD:s andel p_{SD} är större än alla de övriga, dvs.

$$\begin{aligned} H_0 &: \boldsymbol{p} \in \mathbb{S}^{10} - \omega_{SD} \\ H_1 &: \boldsymbol{p} \in \omega_{SD} \end{aligned} \tag{3}$$

Vi föreslår att hypoteserna (3) testas med ett maximum likelihoodkvottest. Detta innebär att vi söker likelihoodens maximum om parameterrummet är begränsat under H_0 och jämför detta med likehoodens maximum om parameterrummet inte är begränsat. Då vi enbart har en observation av en multinomialfördelad slumpvariabel blir likelihooden samma som sannolikhetsfunktionen:

$$L(\boldsymbol{p} \mid \boldsymbol{x}) = \frac{n!}{x_1! \cdots x_{10}!} p_1^{x_1} \cdots p_{10}^{x_{10}} \tag{4}$$

Likelihooden maximeras i det obegränsade parameterrummet av ML-skattningen vilket i detta fall är $\hat{\boldsymbol{p}} = \boldsymbol{x}/n$. I det begränsade parameterrummet under H_0 maximeras (4) av skattningen $\boldsymbol{p}^*$. Om vi antar att antalet SD-sympatisörer i urval x_{SD} är det största värdet i $\boldsymbol{x}$, i annat fall verkar det omotiverat att testa (3), så medför detta att $\boldsymbol{p}^*$ kommer att vara en punkt på randen av $\mathbb{S}^{10} - \omega_{SD}$. Detta innebär att (4) ska maximeras över $\boldsymbol{p}$ under bivillkoren

a) $p_{SD} \leq p_j$, för alla andra partier j,
b) $p_j > 0$, för alla partier j och
c) summan av alla andelar p_j är 1 ($p_1 + \cdots + p_D = 1$).

Tabell 1. De skattade väljarandelarna rapporterade i Metro $\hat{p}$, frekvenser x som skulle motsvara dessa andelar vid OSU, samt skattade andelar om SD inte tillåts vara det största partiet p^*.

Parti	M	C	L	KD	MP	S	V	FI	SD	Övriga
$\hat{p}$	0,210	0,056	0,044	0,037	0,064	0,234	0,068	0,028	0,252	0,007
x	321	85	67	56	98	357	104	43	385	11
p^*	0,210	0,055	0,043	0,036	0,064	0,244	0,070	0,029	0,244	0,007

I normala fall måste p^* skattas numeriskt. De numeriska beräkningarna förenklas avsevärt om de två villkoren b) och c) undanröjs genom att problemet överförs från $\mathbb{S}^{10}$ till $\mathbb{R}^9$ medelst en ILR-transformation. Det första villkoret a) kan då omformuleras som ett antal linjära olikheter $\boldsymbol{uy} \leq \mathbf{0}$. Matrisen $\boldsymbol{u}$ kommer att bero på valet av ILR-transformation och på vilken komponent som antas vara störst, men i vårt fall med tio andelar där SD utgör den nionde så blir

$$\boldsymbol{u} = \begin{bmatrix} -2^{-1/2} & -6^{-1/2} & -12^{-1/2} & -20^{-1/2} & -30^{-1/2} & -42^{-1/2} & -56^{-1/2} & -\sqrt{9/8} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{7/8} & -\sqrt{9/8} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\sqrt{8/9} & \sqrt{10/9} \end{bmatrix}.$$

För att bestämma (4) i en punkt y_0, så måste y_0 först transformeras tillbaka till $p_0 = \mathrm{ILR}^{-1}(y_0)$.

Testfunktionen för maximum likelihoodkvottestet är

$$\lambda = -2\big(\log L(\hat{p}) - \log L(p^*)\big). \tag{5}$$

Om H_0 är sann och p ligger på randen av $\mathbb{S}^{10} - \omega_{SD}$, så kommer (5) att vara 0 med sannolikheten δ, som andelen av sannolikhetsmassan som finns i $\mathbb{S}^{10} - \omega_{SD}$. Sannolikheten δ beror på p, men kommer vanligen att vara omkring 1/2, såvida inte p är nära en punkt där $d \geq 3$ partier är lika stora, i vilket fall δ kommer att vara omkring $(d-1)/d$. Med sannolikhet $1 - \delta$ så kommer (5) vara approximativt χ^2-fördelad med en frihetsgrad. För att bestämma p-värdet för testfunktionen får man således dela sannolikheten att λ överskrider det observerade testvärdet med antal andelar d i p^* som är lika eller ungefär lika stora som den största.

4 Resultat

I Tabell 1 återfinns Metros skattade andelar $\hat{p}$ för de nio redovisade partierna samt andelen för övriga partier bestämd av oss som differensen mellan summan av andelarna och ett. Undersökningen som Metro redovisar är gjord av YouGov. YouGov använder sig av en självrekryterad webpanel och resultaten i detta fall bygger på svar från 1527 respondenter. Detta utgör ju inget slumpmässigt urval, så för att kunna räkna på det så kommer vi att anta att YouGov har använt sig av ett slumpmässigt obundet urval om 1527 individer utan bortfall. Metro presenterar inte hur många av de tillfrågade som har angett respektive parti utan bara en skattning av andelen $\hat{p}$. I Tabell 1 återfinns

observerade frekvenser $\boldsymbol{x}$ bestämda som $n\hat{\boldsymbol{p}}$ lämpligt avrundat. Vi skattar därefter de partiernas andelar $\boldsymbol{p}^*$ om SD inte tillåts vara det största partiet. Även dessa återfinns i Tabell 1.

Det observerade värdet på (5) blir således $-2(-27{,}4522 - (-28{,}0375)) = 1{,}1706$. Sannolikheten att λ ska överskrida 1,1706 är $\Pr(\lambda \geq 1{,}1706) = 0{,}2793$. I skattningen $\boldsymbol{p}^*$ är SD och S lika stora (0,244) men även M är nästan lika stora (0,210), så p-värdet för testet är mellan 0,2793/3 = 0,0931 och 0,2793/2 = 0,1396. Testet ger således inte tillräckligt stöd för att förkasta nollhypotesen. Man kan inte med utifrån YouGovs undersökning dra slutsatsen att SD är det största partiet i väljarkåren i Sverige.

Referenser

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B,* 44, 139-177.

Aitchison, J. (1986). *The statistical analysis of compositional data,* London: Chapman and Hall. (Nytryck med extra material 2003 utgivet av The Blackburn Press.)

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology,* 35, 279-300.

Pawlowsky-Glahn, V., Egozcue, J. J. & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*: John Wiley & Sons

Wallroth, E. 2015. Nu är SD Sveriges största parti. *Metro,* 20 augusti 2015.

Texten bygger i stora delar på Bergman, J. & Holmquist, B. "Are the Sweden Democrats really Sweden's largest party? A maximum likelihood ratio test on the simplex", inskickad för publicering.

Fitting Statistical Models in Time Series Analysis

Paul Fischer
pafi@dtu.dk
DTU Compute
Lyngby

Astrid Hilbert
astrid.hilbert@lnu.se
Linnaeus University
Växjö, Sweden

30. 11. 2015

Abstract

The task of fitting statistical models to time series is one of the fundamental problems in time series analysis. If one wants to fit a single model to the whole series, the most efficient way is to select the fastest known fitting algorithm for the given type of models. For some applications, one has to repeatedly fit models to different parts of the series. In this case a preprocessing of the data can result in a considerable speed-up. We would like to show how, for a number of relevant models, the time for fitting a model to any part of a series can be reduced from at least linear (in the length of the series) to logarithmic. The models include regression, AR-models and MA-models. We have applied our method to a number of real-world problems where a solution would be infeasible without the achieved speed-up.

1 Introduction

Fitting models to time series is one of the fundamental tasks in time series analysis. We address the problem where one has to fit models to parts of the time series, possibly repeatedly. Consider a fixed type of statistical models, e.g., autoregressive models, and a univariate, equidistant time series is a sequence $Y = (y_0, \ldots, y_{T-1})$ of real numbers, where y_t is the observation at time $t \in \{0, \ldots, T-1\}$. Given two indices a, b (the range), $0 \leq a \leq b \leq T-1$, the task is to fit a model of the given type to the sub-series $(y_a, \ldots, y_b)$. For doing this once, one would simply apply a standard method for fitting the chosen type of model. Assuming that the model depends on all observations in the interval this requires at least linear time in the length of the interval $b - a + 1$. For the case that multiple queries with different ranges have to be performed, a preprocessing might pay off. A straightforward preprocessing is to fit models to all possible $T(T+1)/2$ intervals $[a, b]$ in advance and store them in a table. Then a model for a specific interval $[a, b]$ can looked up in constant time. However, the time for this preprocessing is lower bounded by $\Omega(T^2)$. The at least quadratic preprocessing time and, especially, the quadratic space requirement make this approach infeasible already for medium data sizes of around 10,000.

For a number of relevant statistical models we show that with a linear time preprocessing the time to fit a model to a given interval can be reduced to logarithmic in the

length T of the series. For a type of statistical models to be suited for this speed-up they have to satisfy a *merging condition.* Intuitively, this means that given models for two consecutive intervals $[a, c]$ and $[c+1, b]$ on can easily compute a model for the union interval $[a, b]$ for the two. When a a type of model satisfies this condition, then we can use a data structure, a *range tree*, to implement the speed-up.

In Chapter 2 we introduce the framework which allows to speed-up the computations. We use a simple statistical model to exemplify how the methods works and indicate how it can be used for other, more involved models.

2 The Framework

2.1 Range Trees

We start by describing the data structure used for speeding up the model fitting.

Range trees are a data structure which supports multiple *queries* on intervals of indexed data. A description of the general concept of a range tree may be found in [1]. We restrict the presentation to the situation where the data is an univariate time series $(y_0, \ldots, y_{T-1})$, that is, we consider range trees for one-dimensional numerical data. A *range query* receives two indices a, b (the range), $0 \leq a \leq b \leq T-1$, as inputs and returns as *answer* a quantity $q(a, b) = q(y_a, \ldots, y_b)$ which is determined by the data $y_a, \ldots, y_b$. The maximum is one example: $q(a, b) = \max\{y_i \mid a \leq i \leq b\}$.

For a single query on range $[a, b]$, the most efficient way is to compute the answer $q(a, b)$ directly from the data. The worst-case time for a single query is bounded from below by $\Omega(T)$. For the case that multiple queries with different ranges have to be performed, a preprocessing might pay off. A straightforward preprocessing is to compute the answers for all possible $T(T+1)/2$ ranges $[a, b]$ in advance and store them in a table. Then a query $[a, b]$ can be answered by a look-up in the table. The time for the preprocessing is lower bounded by $\Omega(T^2)$. The time for the query is the time to write the stored answer $q(a, b)$. For the example of querying the maximum, the preprocessing time is $\Theta(T^2)$ and the query time is constant. The at least quadratic preprocessing time and, especially, the quadratic space requirement make this approach infeasible already for medium data sizes of around 10,000.

Range trees offer a good compromise. They require the following condition to be met.

Condition 1. a) (Mergability) Given a range $[a, b]$ and an index c, $a < c \leq b$, the answer $q(a, b)$ for the range $[a, b]$ can be computed from the answers $q(a, c-1)$ and $q(c, b)$ for the ranges $[a, c-1]$ and $[c, b]$.

b) (Initialization) The answers $q(a, a)$ for singleton ranges $[a, a]$ can be computed from the data y_a.

Given that a) and b) can be performed in constant time, and any answer can be stored in constant space, the preprocessing time for a range tree is $O(T)$ and the query time is $O(\log(T))$, see [1]. This is the case for our running example of finding the maximum.

A one-dimensional *range tree* is a rooted, binary tree. Every node N covers a range $[a, b]$, i.e., it contains the indices a, b and the answer $q(a, b)$ to the query with that range. If $a \neq b$, the left child of N covers the range $[a, \lfloor (a+b)/2 \rfloor]$, the right child covers the

range $[\lfloor (a+b)/2 \rfloor + 1, b]$. We ignore the technical details for optimally balancing the tree. The root covers the whole range $[0, T-1]$, the leaves cover singleton ranges $[a, a]$.

To answer a range query for the range $[a, b]$, one starts at the root. Then the indices a and b follow the unique paths P_a and P_b to the leaves L_a, L_b containing the ranges $[a, a]$, and $[b, b]$ respectively. Note that the two paths might be identical in the beginning. As soon as the paths have separated we collect and merge the answers of all right children of P_a and all left children of P_b as well as the answers at L_a and L_b. The depth of the tree is $O(\log(T))$. The result of the merging is the answers for the range $[a, b]$, see Figure 1 for an illustration.

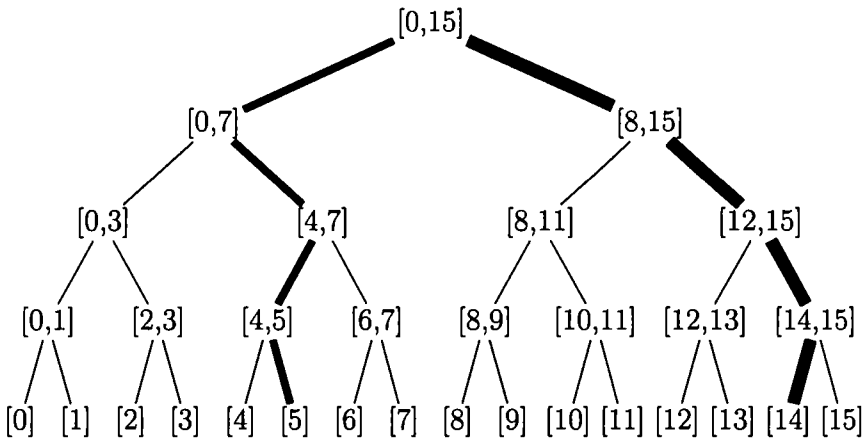


Figure 1: A range tree for 16 data points. For convenience, we denote the leaf nodes by $[a]$ and not by $[a, a]$: For the query range $[5, 14]$ the thick edges indicate the paths P_5 and the even thicker ones P_{14}. The information of the right children of P_5 (node $[6, 7]$) and leaf $L_5 = [5]$ is merged. The information of the left children of P_{14} (nodes $[8, 11]$, $[12, 13]$) and leaf $L_{14} = [14]$ is merged. Finally merging the information for P_5 and P_{14} results in the answer for the range $[5, 14]$.

In order to show how fitting models to parts of a time series can be made fast we introduce a simple, non-parametric statistical model.

2.2 Rectangles

The rectangle model simply is the bounding box R of the subseries from a to b. For a formal definition, let $0 \leq a \leq b \leq T-1$ the model is minimum axes-aligned rectangle which contains the observations $y_a, \ldots, y_b$. That is the rectangle $R(a, b)$ defined by

$$R(a, b) = [a; b] \times [\min\{y_i \mid i = a, \ldots, b\}; \max\{y_i \mid i = a, \ldots, b\}] \,.$$

Given a and b it suffices to know $max(a, b) = \max\{y_i \mid a \leq i \leq b\}$ and likewise the minimum, i.e., the answer to the range query is $q(a, b) = (\min(a, b), \max(a, b))$. In every node of the range tree we store the minimum and maximum for the range covered by the node. Condition 1 for a range tree to be applicable are trivially met: For the merging

property, given intervals $[a; c-1]$ and $[c; b]$ we have

$$\max(a, b) = \max\{\max(a, c-1)\,,\, \max(c, b)\} \tag{2.1}$$

and analogously for the minimum. For the initialization of a singleton range $[a; a]$ the minimum and maximum are both equal to y_a. Thus

Theorem 2.1. *Given a time series of length T, a minimum axis-aligned rectangle can be fitted to any interval $[a; b]$, $0 \le a \le b \le T-1$, in time $O(log(T))$ after a preprocessing which uses $O(T)$ time and space.*

Proof. In the range tree the information of at most $2\log_2(T)$ nodes has to merged. Equation (2.1) each merge operation can be performed in constant time.

The values at the leaf nodes of the tree can be computed in constant time. The values at the interior nodes are computed by merging the values of the two children. As there are $T-1$ interior nodes and merging can be performed in constant time, the preprocessing can be done in $O(T)$. Each nodes holds 2 values thus the space requirement is $O(T)$. □

2.3 Other Models

For the following statistical models the same speed-up as for rectangles can be achieved.

- Linear and polynomial regression.
- Auto-regressive models of arbitrary order p (AR(p)-models).
- Moving average models of order 1 (MA(1)-models).

The data stored in the nodes of the range is more complex for these models than for the rectangle model, but satisfies the merging property. For example, AR(p)-models can be fitted by solving the Yule-Walker equations. To set up these equations, one has to computed the autocovariance coefficients for lags $0, 1, \ldots, p$. In the node of the range tree which covers the interval $[a, b]$ we store the following values

$$S_\ell = \sum_{i=a}^{b} y_{i+\ell}\,, \qquad Q_\ell = \sum_{i=a}^{b} y_i y_{i+\ell}\,, \qquad \text{for } \ell = 0, 1, \ldots, p\,.$$

From those one can compute the autocovariance r_l for lag ℓ by

$$r_\ell = \frac{Q_\ell}{(b-a-\ell+1)} - \frac{S_0 S_\ell}{(b-a-\ell+1)^2}\,. \tag{2.2}$$

Given consecutive intervals $[a, b]$ and $[b+1, c]$ the S and Q values for the interval $[a, c]$ is the sum of the respective values for the two intervals. Thus mergebility holds.

2.4 Running times

Using range trees instead of direct computations imposes some overhead, especially due to the preprocessing. On the other hand the time to fit the models (which means a query to the range tree) is much less.

The tests described here have been performed on artificially generated time series which consist of three regimes each of which is generated by an AR(3) model. The series have lengths $T = 400$, $T = 50,000$, and $T = 1,000,000$. The break-even point, i.e., the number of queries needed to compensate for the preprocessing becomes lesser for longer series. When using a range tree, about 400 queries are enough for the preprocessing to pay off and each fit is about 40 respectively 500 times faster for $T = 50,000$ and $T = 1,000,000$, respectively. When one allows query intervals only to start and end at every 10-th index, the break-even point is already reached at 40 queries for the longer two series and the speed up per fit is improved by a factor of about 50 and 600, respectively. The results are shown in Tables 1 and 2.

For the test shown in the two tables we employed a generic Java implementation, which uses a great deal of object-oriented paradigms. The reference method computes the autocovariances directly from the series, and is a fast, quick-and-dirty implementation. A non-generic (i.e. specialised for either rectangles or AR-models) and quick-and-dirty implementation of the range tree method in Java gives gives another speed-up by a factor 5. For the queries, we selected a, b at random and queried the interval $[a, b]$ and an AR-model was fitted to $[a, b]$. The tables show averages over 10,000 queries.

T	Nodes	build	query	qry.& fit	direct fit	gain	break-even
400	785	1 479	0.96	2.93	2.88	0.98	∞
50 000	99 985	49 524	2.61	3.29	131.18	39.89	388.0
1 000 000	1 999 985	988 216	3.64	5.31	2 719,30	512.27	365.0

Table 1: Average running times in microseconds (μs) for the range-tree data structure for AR-models. Used was a single-threaded implementation in Java, run on a computer with an Intel Core i7-2600 at 3.40 Ghz and 16GB ram. In addition the gain factor per query-and-fit is shown, as is the number of query-and-fit operations needed to compensate the preprocessing time. Shown are the length of the series, number of nodes in the range tree, time to build the range tree, time for a random query, time for a random query plus fitting an AR-model to the query range, and the time to fit an AR-model to a random interval by directly computing the parameters form the series. The last two columns show the speed-up factor by using the range tree (per query and fit) and the number queries needed to compensate the cost for building the range tree data structure.

3 Applications

We present two application areas for the aforementioned techniques.

In econometrics one often assumes that the mechanism which generates the time series can be well modelled by statistical processes. Examples for such processes are MA-, AR-, ARMA-, ARIMA-, GARCH-models etc. All these models are parametric. An important

T	Nodes	build	query	qry.& fit	direct fit	gain	break-even
400	77	1 180	0.85	2.67	2.92	1.09	4742.0
50 000	9 997	3 883	1.83	2.84	136.73	48.08	30.0
1 000 000	199 997	108 891	3.98	4.34	2 737,83	630.53	40.0

Table 2: Average running times in microseconds (μs) for the range-tree data structure for AR-models. Setup as in Table 1 with the difference that the range tree here uses a minimum interval length of 10 (interval boundaries only allowed at indices $0, 10, 20, \ldots$). To use the range tree data structure with interval length 10 asymptotically pays off for series of length $T \geq 400$, practically for $T \geq 800$.

Figure 2: The plot shows the development of the parameters ϕ_1 (top graph) and ϕ_2 (bottom graph) from Equation (3.1). The original data are the closing quotes form November 1990 to Ferbruary 2012, in all 5372. To compensate for drift, logarithms were taken. Then, to remove seasonal effects, a spline with 22 knots was fitted and removed. The window size is $w = 200$ and the window was moved by 10 days in each step, resulting 518 models to be fitted.

issue for the analysis is, whether the parameters of the undelying model change with time. In order to detect gradual changes or to monitor the development of the model, one can take the following approach. Given a time series of length T, a sliding window of length $w < T$ is moved along the series. At each position of the window, a model of a fixed type is fitted to the part of series inside the window. The parameters of each model are recorded to monitor the development. Figure 2 shows an example for fitting an AR(2)-model to the daily closings of the German stock index DAX. An AR(2) model predicts the next observation y_i by

$$y_i = c + \phi_1 y_{i-1} + \phi_2 y_{i-2} + \varepsilon_i \tag{3.1}$$

where c is a constant, and the random shock ε_i is $\mathcal{N}(0, \sigma)$ distributed for some $\sigma > 0$.

The second application is the detection of break points (also called change points) in a time series. A break indicates a significant change in the behaviour of the series. One method to formalize the notion of a break point, is to fit statistical models piecewise to the series. A structural break is indicated by a significant change of the model parameters in adjacent pieces. To detect such changes, one varies the pieces and repeatedly fits models to them. This is usually computationally very expensive, because many models have to be fit to different parts of the series.

One selects the type of statistical model to be used and one (or more) set(s) of candidate

break points. Then models of the given type are fit to each of the intervals between consecutive break points. For each of the models it is evaluated how good it represents the data and an overall fit-score is computed. Then some candidate points are (randomly) moved (or removed or new ones are added), models are fit and a new fit score is computed. If the new fit-score is better than the previous one, then one assumes that the new candidate points are closer to true break points than the previous ones. In this case one continues with new set of candidate points, otherwise one uses the old one again. The process is terminated, when a stop criterion is met, e.g., no improvements occurred in the last k rounds.

Moving, removing or adding new candidate points can be done manually, by a deterministic algorithm or a randomized search heuristic. Examples for the latter, using a evolutionary algorithm, can be found in [2, 3]. Depending on the length of the series and the desired accuracy between 10,000 and 200,000 models have to be fitted.

We applied out method in an industrial application where different working conditions of a machine had to be identified from measurements of vibrations. The series consisted of up to 1,000,000 observations. The classification made by our algorithms is in very good accordance with the one of the engineers. All changes of the different working states of the machine were found as break points by our application. For this application the specific statistical model described in Section 2.2 is essential because it allows the aforementioned speed-up of the algorithm. Otherwise it would not be possible to process the long time series of around one million observations in acceptable time, see Figure 3.

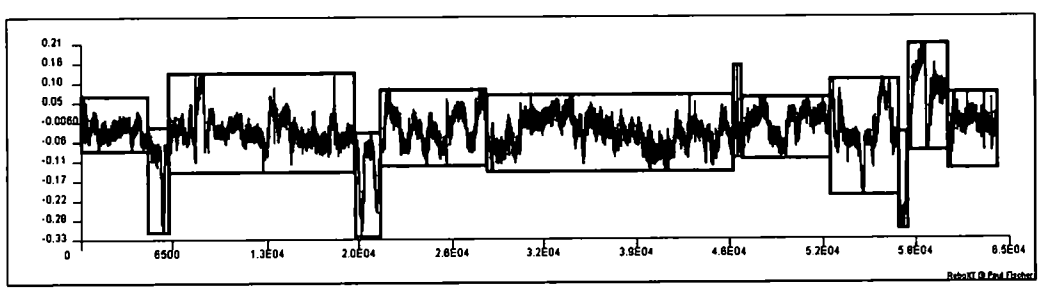


Figure 3: The result of the structural break detection for a time series of vibration patterns. The regimes coincide nicely with the different working states that were recorded while the machine was working.

References

[1] M. de Berg, M. van Krefeld, M. Overmars, O. Schwarzkopf, Computational Geometry: Algorithms and Applications, 2nd Edition, Springer, 2000.

[2] B. Doerr, P. Fischer, A. Hilbert, C. Witt, Evolutionary algorithms for the detection of structural breaks in time series, in: Proceedings of the 15th annual conference companion on Genetic and evolutionary computation, ACM, 2013, pp. 119–120.

[3] P. Fischer, A. Hilbert, Fast detection of structural breaks, in: Proceedings of 21th International Conference on Computational Statistics 2014, 2014, pp. 9–16.

Multivariate Time Series Estimation using marima

Henrik Spliid, DTU Compute

A computer program, called `marima`, written in the `open source` language, *R*, has been developed. Some of `marima`'s facilities and ideas are presented in the following.

1 The multivariate ARMA(p,q) model

Let t denote (discrete) time. Consider a k-variate random vector y_t of *observations* and, correspondingly, a k-variate random vector, u_t, of *unknown innovations* with zero mean:

$$y_t = \begin{pmatrix} y_{1,t} \\ y_{2,t} \\ \cdots \\ y_{k,t} \end{pmatrix} \quad , \text{ and } \quad u_t = \begin{pmatrix} u_{1,t} \\ u_{2,t} \\ \cdots \\ u_{k,t} \end{pmatrix} \quad , \quad t = \{1, 2, \ldots, n\} \tag{1}$$

Further suppose that the random vector y_t is generated through the model

$$y_t + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} = u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} \tag{2}$$

where the coefficient matrices $\phi_1, \ldots, \phi_p$ and $\theta_1, \ldots, \theta_q$ are all of dimension $k \times k$.

The series u_t is without autocorrelation, but the individual elements (coordinates) need not be, for example, uncorrelated. The covariance matrix of u_t is $\mathrm{Var}(u_t) = \Sigma_u$, and it is assumed to be independent of t. Most often u_t is assumed to be normally distributed.

The lefthand side of equation (2) is called the *autoregressive* or AR part of the model, while the righthand side is called the *moving average* or MA part of the model. p is the order of the AR part, and q is the order of the MA part. The model is called the *multivariate* $ARMA(p, q)$ model.

The model is often extended to include external non-random regression variables.

A further and somewhat more detailed description of `marima` is available from the repository where `marima` is located (contact: hspl[at]dtu.dk) .

Henrik Spliid, professor emer.
DTU, Bygning 324,
Danmarks Tekniske Universitet
2800 Lyngby

http://www.compute.dtu.dk
Telefon : +45 4525 3362
E-mail :hspl[at]dtu.dk

2 Operator form of the ARMA(p,q) model

2.1 Matrix polynomials used in marima

Define the *k-variate backwards shift operator* B such that if B is multiplied on a time indexed k-variate random variable, the result is to be interpreted as the variable lagged a single time step. And, in general, lagging r time steps is accomplished using:

$$B^r y_t = y_{t-r}$$

Now, introduce the operator B into model (2) which then can be written as

$$(I + \phi_1 B + \cdots + \phi_p B^p) y_t = (I + \theta_1 B + \cdots + \theta_q B^q) u_t \ ,$$

where I is the $k \times k$ unity matrix. Also, define the *matrix polynomials* $\phi(B) = I + \phi_1 B + \cdots + \phi_p B^p$ and $\theta(B) = I + \theta_1 B + \cdots + \theta_q B^q$. This leads to the general multivariate ARMA(p,q) model in operator form

$$\phi(B) y_t = \theta(B) u_t \ , \tag{3}$$

Most often the averages of the k variables in y_t are subtracted before the model estimation. When reconstructing or forecasting the measured series analysed by `marima`, the averages of the original data can be reintroduced.

2.2 Averages and their representation in the arma model

Suppose that the vector y_t has been (for example) means-adjusted, such that $y_t = v_t - \eta$, where v_t are the original measurements, and η is the (estimated) vector of averages: $\eta \simeq E\{v_t\}$. Suppose now, that $\phi(B) y_t = \theta(B) u_t$ or, equivalently:

$$\phi(B)(v_t - \eta) = \theta(B) u_t$$

$$\phi(B) v_t = \mu + \theta(B) u_t \quad ; \quad \mu = \phi(B)\eta$$

$$\mu = \left[\sum_{i=0}^{p} \phi_i \right] \eta \tag{4}$$

Note that equation (4) applies for any transformation of the form $y_t = z_t - \eta$.

2.3 Inverse matrix polynomials and alternative forms

It is convenient to be able to write model (3) in the following form:

$$y_t = u_t + \psi_1 u_{t-1} + \cdots + \psi_\ell u_{t-\ell} + \cdots = \psi(B)u_t \tag{5}$$

Given the model (3), the model (5) can be determined if we are able to calculate the *left inverse polynomial* $\phi^{-1}(B)$ of the polynomial $\phi(B)$, such that

$$\phi^{-1}(B)\phi(B) = I \tag{6}$$

In general, if $\phi(B)$ is a finite order polynomial, the inverse, $\phi^{-1}(B)$ is of infinite order.

Pre-multiplying with $\phi^{-1}(B)$ on both sides of the equals sign in model (3) gives

$$y_t = \phi^{-1}(B)\theta(B)u_t = \psi(B)u_t = \sum_{i=0}^{\infty} \psi_i u_{t-i} \tag{7}$$

This form is called the *random shock* form and, generally (if the model includes a nonzero AR term), the polynomial $\psi(B)$ is of infinite length with decreasing coefficients, such that $\psi(\ell) \to 0$ for $\ell \to \infty$. If, more precisely, $\sum_{i=0}^{\infty} \psi(z)$ converges for all $|z| \leq 1$ the model (7) is said to be *stationary.*

Similarly if $\theta^{-1}(B)$ is the left inverse of $\theta(B)$, we may pre-multiply with $\theta^{-1}(B)$ on both sides of the equals sign in model (3). This gives the socalled *inverse form:*

$$\pi(B)y_t = u_t \tag{8}$$

where $\pi(B) = \theta^{-1}(B)\phi(B)$. If, similarly, $\sum_{i=0}^{\infty} \phi(z)$ converges for all $|z| \leq 1$ the model (8) is said to be *invertible.*

3 Simple operations for matrix polynomials

Note that, the 0'th order coefficient matrix of $\phi(B)$ (that is ϕ_1) and of $\theta(B)$ is the $k \times k$ unity matrix.

The *marima* package includes routines for inverting and multiplying matrix polynomials, namely `pol.inv` and `pol.mul`.

The left inverse of `phi` is computed as, say, `inv.phi <- pol.inv(phi,L)` which will result in an array, `inv.phi`, of dimension $k \times k \times (1 + L)$ holding the $k \times k$ unity matrix followed by the first `L` matrix coefficients of the left inverse of `phi`.

The product of two matrix polynomials, $\phi(B)$ and $\theta(B)$, is computed as `pol.mul(phi,theta,L)` which will result in an array of dimension $k \times k \times (1+L)$ holding the $k \times k$ unity matrix followed by the first `L` matrix coefficients of the product $\phi(B)\theta(B)$.

Equation (7) can be carried out using, for example, `psi<-pol.mul(pol.inv(phi,L), theta,L)` giving the unity matrix of order $k \times k$ followed by the first `L` matrix terms of the $\psi(.)$ polynomium. The array `psi` will have dimension $k \times k \times (1+L)$.

4 Differencing multivariate time series

Often it is convenient to do differencing, such as $z_t = y_t - y_{t-s}$, that is differencing over s time steps. A routine called `define.dif` included in the marima package can perform (practically) all kinds of differencing for a multivariate time series.

4.1 Single time step differencing

The polynomial $\nabla(B) = (I - B)$ can be used to difference the time series one time step, for example

$$z_t = \nabla(B)y_t = y_t - y_{t-1} \tag{9}$$

If, for example, $y_t = \{y_{1,t}, y_{2,t}\}^T$ is bivariate then the ∇ polynomial for differencing both variables once is

$$\nabla(B) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix}$$

4.2 Seasonal differencing

The polynomial $\nabla(B^s) = (I - B^s)$ is used if a seasonal differencing with seasonality s is wanted:

$$z_t = \nabla(B^s)y_t = y_t - y_{t-s} \tag{10}$$

The polynomial for s timesteps seasonal differencing is

$$\nabla(B^s) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} B^s & 0 \\ 0 & B^s \end{pmatrix}$$

4.3 Mixed differencing

When a multivariate time series is at hand, it may be necessary to difference the individual series differently.

Suppose again, that $y_t = \{y_{1,t}, y_{2,t}\}^T$ is bivariate and that we want to difference over time periods $s = \{s_1, s_2\}$. Then we may define, a little more generally,

$$\nabla(B^s) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} B^{s_1} & 0 \\ 0 & B^{s_2} \end{pmatrix}$$

and

$$\nabla(B^s) \begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} = \begin{pmatrix} y_{1,t} \\ y_{2,t} \end{pmatrix} - \begin{pmatrix} y_{1,t-s_1} \\ y_{2,t-s_2} \end{pmatrix}$$

The routine `define.dif(...)` can perform mixed differencing of a multivariate series. The routine `define.sum(...)` included in the marima package does the reverse, that is summing a multivariate series.

4.4 Aggregated model

Suppose the differencing polynomium used (before analysing the data by `marima`), is

$$\nabla(B) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} B - \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} B^{12}$$

The differenced series is called z_t, and $z_t = \nabla(B)y_t$. Suppose now that z_t is analysed by `marima` and that the estimated model for z_t is $\widehat{\phi}(B)z_t = \widehat{\theta}(B)u_t$. The estimated aggregated (nonstationary) model for the observed time series, y_t, is then $\widehat{\phi}(B)\nabla(B)y_t = \widehat{\theta}(B)u_t$.

5 Model selection in marima

5.1 Defining the marima model

Defining models in `marima` is done by creating 0/1-indicator arrays corresponding to the ar-part and the ma-part of the model. These arrays are organised the same way as the model polynomiums wanted. Suppose the ar-indicator array is called `ar.pattern`. Then the value at position `ar.pattern[i,j,ℓ]` is the indicator for the $\{i,j\}$'th element in the lag=ℓ ar-parameter matrix $\phi_\ell = \{\phi_{i,j}\}_\ell$.

The value 1 (one) indicates that a parameter is to be estimated at that position. The value 0 indicates that the parameter corresponding to that position is 0. The function `define.model` can be used for setting up the indicator arrays properly.

Examples of the use of `define.model` are obtained with
`library(marima)`
`example(define.model)`.

6 Estimation and identification of a marima model

As described by Spliid (1983) the estimation of the model is done with a pseudo-regression method. In the present implementation the `R` procedure `lm(...)` is used. This enables `marima` to utilize the procedure `step(...)` in order to search for a good model in a stepwise manner. The key parameter in `step` is the k-factor used in Akaike's criterion (where k=2 is Akaike's suggestion).

The k-factor can be set when calling `marima` by specifying the input parameter `penalty` to an alternative value.

After having estimated the `marima`-model as defined with the use of, for example, `define.model`, the value `penalty=0` causes `marima` to do no search for a (reduced) model. If `penalty=2` the usual AIC is used to identify a reduced model and in a stepwise manner.

This is repeated a few times and from then on, `marima` iterates on the selected model until (hopefully) convergence.

Experience has shown that a first choice `penalty=1` often results in a model which gives a good overview of which coefficients are the most important ones and which are less important. Approximate F-test values for the individual parameter estimates are given in the output object, and they serve the same purpose.

7 Australian firearms legislation

Baker & McPhedran (2007) discuss the effect of the Australian firearms legislation of (implemented) 1997 on death rates. Four different (maybe related?) death rates (firearm suicides, firearm homicides, other suicides and other homicides) are considered. Baker & McPhedran analysed the data by conventional univariate ARIMA models with separate models for each of the four death rates. All models estimated were univariate arma(1,1) models, i.e. arma models of order (ar=1,ma=1) and without differerencing.

Here, we shall illustrate the use af `marima` along the same lines, although it is by no means claimed that the results obtained are optimal or represent the best analysis of these data.

7.1 Data

The data for the study can be accessed using, for example,
`library(marima); data(australian.killings); all.data <- austr` .

The data.frame `austr` has the following appearance, in that the last 10 lines correspond to not observed future values:

	Year	suic.fire	homi.fire	suic.other	homi.other	leg	acc.leg
1	1915	4.031636	0.5215052	9.166456	1.303763	0	0
2	1916	3.702076	0.4248284	7.970589	1.416094	0	0
3	1917	3.056176	0.4250311	7.104091	1.052458	0	0
4	1918	3.280707	0.4771938	6.621064	1.312283	0	0
5	1919	2.984728	0.8280212	7.529215	1.309429	0	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
80	1994	2.4027240	0.2744370	9.297252	1.3385800	0	0
81	1995	2.2023310	0.3209428	10.397440	1.4829770	0	0
82	1996	2.1025940	0.5406671	10.900720	1.1632530	1	1
83	1997	1.7982930	0.4050209	12.901270	1.3284680	1	2
84	1998	1.2024840	0.2885961	13.099060	1.2345500	1	3
85	1999	1.4002010	0.3275942	11.698280	1.4847410	1	4
86	2000	1.2008320	0.3132606	11.099870	1.3365790	1	5
87	2001	1.2980830	0.2575562	11.301570	1.3392920	1	6
88	2002	1.0997420	0.2138386	10.702110	1.4052250	1	7
89	2003	1.0013770	0.1861856	10.099310	1.3334910	1	8
90	2004	0.8361743	0.1592713	9.591119	1.1497400	1	9
91	2005	NA	NA	NA	NA	1	10
92	2006	NA	NA	NA	NA	1	11
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
100	2014	NA	NA	NA	NA	1	19

It is noted that the data.frame is organised columnwise, while the time series in principal should be organised rowwise. This is (if needed) taken care of in `marima` such that the datamatrix is transposed if the number of rows is larger than the number of columns.

The column `leg` indicates whether legislation has been imposed or not (1 or 0). The column `acc.leg` accumulates the legislation (as one possible parametrisation of the effect of the legislation).

Note, that `leg` is set to 1 already in 1996. This is because the first effect of `leg` will be for the year *after* 1996 (namely 1997). This is a general feature in time series models where present values depend on previous values. The first year where `leg` can (is believed to) have an effect is therefore 1997.

7.2 Analysis of the four-variate time series

We will estimate the four univariate the models for the four death rates for the period from 1915 to 1996 (both included) as discussed by Baker & McPhedran. In order to define the model the procedure `define.model` is used, and subsequently `marima` is called using the data from the period 1915 to 1996:

```
rm(list=ls())
library(marima)
data(australian.killings)
old.data <- t(austr)[,1:83]
ar<-c(1)
ma<-c(1)
# Define the proper model:
Model1   <- define.model(kvar=7, ar=ar, ma=ma, rem.var=c(1,6,7), indep=c(2:5))
# Now call marima:
Marima1  <- marima(old.data,means=1,
 ar.pattern=Model1$ar.pattern, ma.pattern=Model1$ma.pattern,
             Check=FALSE, Plot=FALSE, penalty=0.0)
short.form(Marima1$ar.estimates, leading=FALSE) # print estimates
short.form(Marima1$ma.estimates, leading=FALSE)
```

Using `define.model` the variables in the data which are irrelevant for the analyses are taken out, `rem.var=c(1,6,7)`, and `indep=c(2:5)` results in the variables 2, 3, 4 and 5 being analysed independently. The estimated model is as follows:

```
 > short.form(Marima1$ar.estimates, leading=FALSE)
, , Lag=0 (unity matrix, not printed here (leading=FALSE))

, , Lag=1

   x1=y1  x2=y2  x3=y3  x4=y4  x5=y5 x6=y6 x7=y7
y1     0  0.0    0.0    0.0    0.0     0     0
y2     0 -0.7932 0.0    0.0    0.0     0     0
y3     0  0.0   -0.7848 0.0    0.0     0     0
y4     0  0.0    0.0   -0.8870 0.0     0     0
y5     0  0.0    0.0    0.0   -0.9809  0     0
y6     0  0.0    0.0    0.0    0.0     0     0
y7     0  0.0    0.0    0.0    0.0     0     0
```

```
> short.form(Marima1$ma.estimates, leading=FALSE)
, , Lag=0 (unity matrix, not printed here (leading=FALSE))

, , Lag=1

   x1=y1   x2=y2   x3=y3  x4=y4   x5=y5 x6=y6 x7=y7
y1     0  0.0     0.0    0.0     0.0        0     0
y2     0 -0.1317  0.0    0.0     0.0        0     0
y3     0  0.0    -0.4162 0.0     0.0        0     0
y4     0  0.0     0.0    0.0163  0.0        0     0
y5     0  0.0     0.0    0.0    -0.7104     0     0
y6     0  0.0     0.0    0.0     0.0        0     0
y7     0  0.0     0.0    0.0     0.0        0     0
```

Further statistics are saved in the object `Marima1`. For example the covariance matrix of the residuals (`Marima1$resid.cov`):

```
> round(Marima1$resid.cov[2:5,2:5], 4)
       u2     u3      u4      u5
u2 0.1083 0.0161  0.0354  0.0112
u3 0.0161 0.0146  0.0041  0.0020
u4 0.0354 0.0041  0.6582 -0.0041
u5 0.0112 0.0020 -0.0041  0.0224
```

and the covariance matrix of the original variables (`Marima1$data.cov`):

```
> round(Marima1$data.cov[2:5,2:5], 4)
       y2     y3     y4     y5
y2 0.2348 0.0425 0.1110 0.0296
y3 0.0425 0.0199 0.0552 0.0097
y4 0.1110 0.0552 2.3522 0.0790
y5 0.0296 0.0097 0.0790 0.0573
```

The multiple correlations for the 4 variables are:

```
> round(1-(diag(Marima1$resid.cov[2:5,2:5])/diag(Marima1$data.cov[2:5,2:5])),2)
  u2   u3   u4   u5
0.54 0.27 0.72 0.61
```

We may now estimate the general four-variate arma(1,1) model. The only modification in comparison with the above analysis is the model (Model2) definition statement where `indep=c(2:5)` is taken out (`indep=NULL`).

```
ar<-c(1)
ma<-c(1)
Model2   <- define.model(kvar=7, ar=ar, ma=ma, rem.var=c(1,6,7), indep=NULL)
Marima2  <- marima(old.data, means=1, ar.pattern=Model2$ar.pattern,
     ma.pattern=Model2$ma.pattern, Check=FALSE, Plot=FALSE, penalty=0)
```

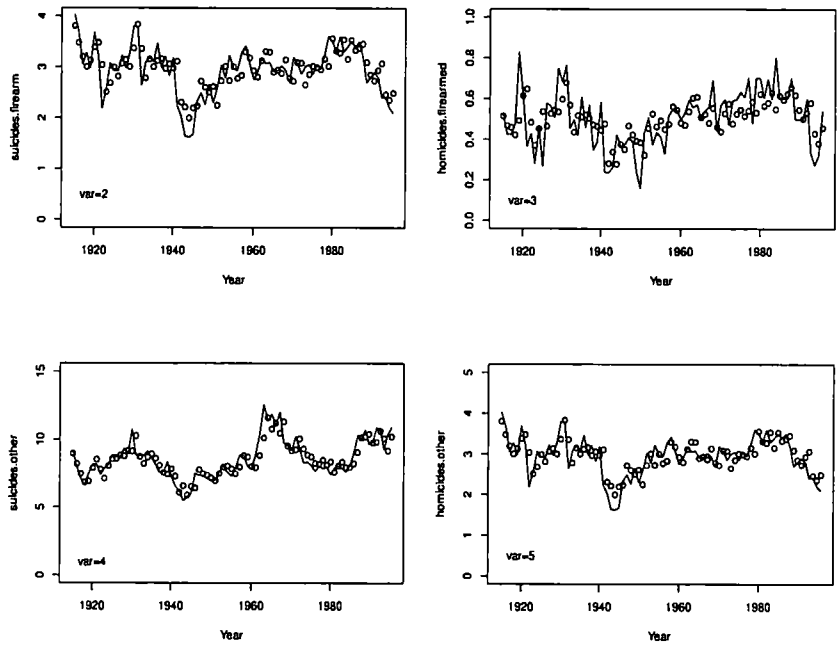
```
> short.form(Marima2$ar.estimates,leading=FALSE)
, , Lag=1

   x1=y1   x2=y2   x3=y3   x4=y4   x5=y5 x6=y6 x7=y7
y1     0  0.0     0.0     0.0     0.0        0     0
y2     0 -0.5916 -0.8260  0.0545 -0.0583     0     0
y3     0 -0.0933 -0.0868 -0.0034 -0.0861     0     0
y4     0  1.2875 -4.1046 -0.8282 -0.6410     0     0
y5     0  0.1222 -0.6610  0.0164 -0.9458     0     0
y6     0  0.0     0.0     0.0     0.0        0     0
y7     0  0.0     0.0     0.0     0.0        0     0

> short.form(Marima2$ma.estimates,leading=FALSE)
, , Lag=1

   x1=y1   x2=y2   x3=y3  x4=y4   x5=y5 x6=y6 x7=y7
y1     0  0.0     0.0    0.0     0.0        0     0
y2     0  0.0378 -0.1077 0.1750 -0.2258     0     0
y3     0 -0.0146  0.2216 0.0415 -0.0097     0     0
y4     0  1.0700 -2.3897 0.0523 -0.2860     0     0
y5     0  0.0885 -0.5220 0.0308 -0.6557     0     0
y6     0  0.0     0.0    0.0     0.0        0     0
y7     0  0.0     0.0    0.0     0.0        0     0
```

In order to evaluate the improvement in taking into account the correlations between the four variables we may compute:

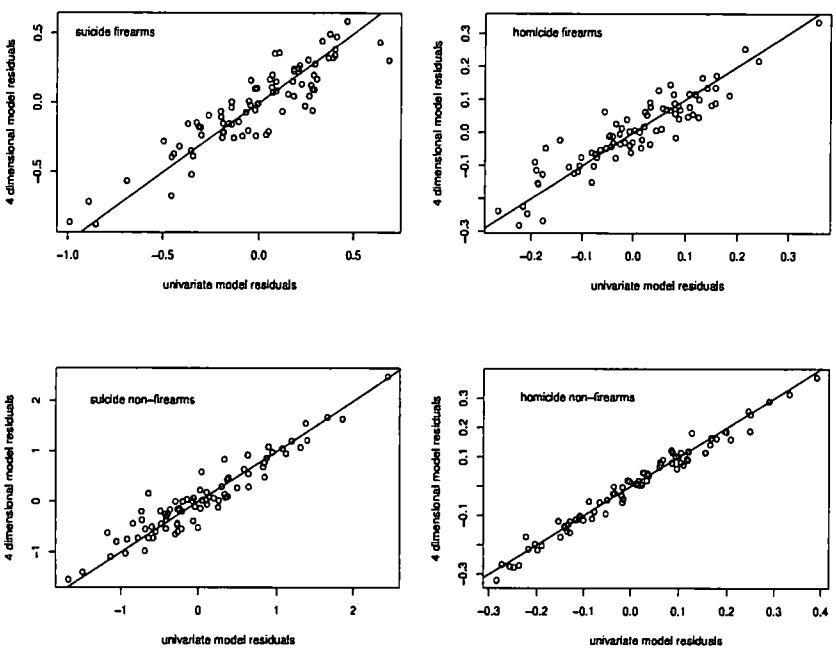
```
>  round(diag(Marima2$resid.cov/Marima1$resid.cov)[2:5], 2)
  u2   u3   u4   u5
0.84 0.89 0.85 1.00
```

so that, for example, the residual variance of the predictions for the first variable (suicides by firearms) estimated by the 4-dimensional model is (only) 84% of the corresponding residual variance for the 4-independent variables model. For the fourth variable (homicides by firearms) there is practically no improvement using the 4-dimensional model.

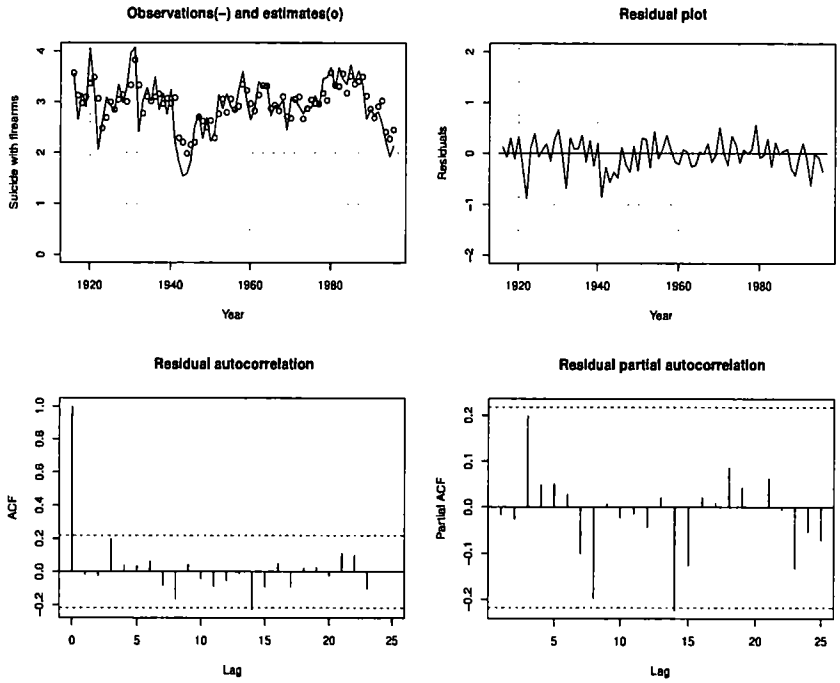
The observations and the predictions for all four variables are shown below (lines=predictions, points=data):



A comparison of the residuals from the univariate models and the 4-dimensional was performed. It turned out that the major differences were for the variable no. 2 (suicides using firearms):



Below are shown the usual model control plots for the 'suicides using firearms' data from the four-dimensional model:



The residual autocorrelations and the partial autocorrelations indicate that the model fits the data adequately.

7.3 Estimation of legislation effect

We shall now estimate a regression model in which variables 6 and 7 are acting as regression variables (use `reg.var=c(6,7)` when calling the model definition procedure `define.model`).

7.3.1 Multivariate model with legislation regression

```
library(marima)
data(australian.killings)
all.data<-t(austr)[,1:90]
ar<-c(1)
ma<-c(1)
Model3   <- define.model(kvar=7, ar=ar, ma=ma, rem.var=c(1), reg.var=c(6,7))
Marima3  <- marima(all.data,means=1, ar.pattern=Model3$ar.pattern,
```

```
    ma.pattern=Model3$ma.pattern, Check=FALSE, Plot=FALSE, penalty=0)
, , Lag=1

> short.form(Marima3$ar.estimates,leading=FALSE)
   x1=y1   x2=y2   x3=y3   x4=y4   x5=y5   x6=y6   x7=y7
y1     0  0.0     0.0     0.0     0.0     0.0     0.0
y2     0 -0.3922 -1.6062  0.0532 -0.0742  0.7155 -0.0163
y3     0 -0.0569 -0.2544 -0.0024 -0.0812  0.0773  0.0107
y4     0  0.8260 -2.7354 -0.7853 -0.5335 -0.9404  0.3087
y5     0  0.1167 -0.6289  0.0140 -0.9501 -0.0257  0.0084
y6     0  0.0     0.0     0.0     0.0     0.0     0.0
y7     0  0.0     0.0     0.0     0.0     0.0     0.0

> short.form(Marima3$ma.estimates,leading=FALSE)
, , Lag=1

   x1=y1  x2=y2   x3=y3  x4=y4   x5=y5 x6=y6 x7=y7
y1     0 0.0     0.0    0.0     0.0        0     0
y2     0 0.2052 -0.8038 0.1557 -0.3363     0     0
y3     0 0.0176  0.0605 0.0374 -0.0181     0     0
y4     0 0.7415 -1.0624 0.0903 -0.0894     0     0
y5     0 0.0879 -0.4930 0.0221 -0.6900     0     0
y6     0 0.0     0.0    0.0     0.0        0     0
y7     0 0.0     0.0    0.0     0.0        0     0
```

One can assess the significance of the estimated coefficients by means of the `Marima3$ar.fvalues` and the `Marima3$ma.fvalues`.

7.3.2 Multivariate model with legislation regression, 'penalty' reduced

The model reduction/simplification is performed using the option 'penalty', for example `penalty=1`. With the previous example but now penalty=1 we get:

```
Marima4  <- marima(all.data, means=1, ar.pattern=Model4$ar.pattern,
     ma.pattern=Model4$ma.pattern, Check=FALSE, Plot=FALSE, penalty=1)


round(short.form(Marima4$ar.estimates,leading=FALSE),4)
, , Lag=1

   x1=y1   x2=y2   x3=y3   x4=y4   x5=y5   x6=y6  x7=y7
y1     0  0.0000  0.0000  0.0000  0.0000  0.0000 0.0000
y2     0 -0.5619 -0.8153  0.0342  0.0000  0.5733 0.0000 (suicide with f.a.)
```

```
y3     0 -0.0608 -0.3171  0.0000 -0.0669  0.0966 0.0000 (homicide with f.a.)
y4     0  0.6533 -1.6631 -0.8268 -0.4720 -0.9667 0.3253 (suicide without f.a.)
y5     0  0.0000  0.0000  0.0000 -0.9801  0.0000 0.0000 (homicide without f.a.)
y6     0  0.0000  0.0000  0.0000  0.0000  0.0000 0.0000
y7     0  0.0000  0.0000  0.0000  0.0000  0.0000 0.0000
```

```
> round(short.form(Marima4$ma.estimates,leading=FALSE),4)
, , Lag=1

   x1=y1  x2=y2 x3=y3  x4=y4   x5=y5 x6=y6 x7=y7
y1     0 0.0000     0 0.0000  0.0000     0     0
y2     0 0.0000     0 0.1303 -0.2351     0     0
y3     0 0.0000     0 0.0391  0.0000     0     0
y4     0 0.5857     0 0.0000  0.0000     0     0
y5     0 0.0000     0 0.0000 -0.7163     0     0
y6     0 0.0000     0 0.0000  0.0000     0     0
y7     0 0.0000     0 0.0000  0.0000     0     0
```

It is seen that many of the regression coefficients for the intervention (x6) and the regression (x7) in the *penalty=1* reduced model are 0 (zero). For variable y2 (suicides with firearms) a constant *decrease* of 0.5733 and no annual *decrease or increase* from 1997 and onwards is found. For variable 3 (homicide with firearms) a small constant *increase* of 0.0966 and practically no annual *change* is found. For variable 4 (suicide without use of firearms) a constant *increase* of 0.9667 and an annual *decrease* of 0.3253 per year is found, but no change of level. For variable 5 (homicide without use of firearms) no effect from the legislation is found.

One might conclude that the level of the rate of suicides using firearms is decreased by about 0.5733 with no annual effect. But suicides without using firearms decreases by about 0.3253 per year after an initial increase of about 0.9667. The rate of homicides (with or without the use of firearms) is generally not affected by the legislation.

In order to assess the significance of the model found one may use the F-values of the ar-part of the estimated model:

```
> round(short.form(Marima4$ar.fvalues, leading=FALSE), 2)
, , Lag=1

   x1=y1 x2=y2 x3=y3  x4=y4  x5=y5 x6=y6 x7=y7
y1     0  0.00  0.00   0.00   0.00  0.00  0.00
y2     0 36.96  6.79   1.43   0.00  8.53  0.00
y3     0  3.04  7.54   0.00   1.44  2.14  0.00
y4     0  5.11  4.55 181.42   1.59  1.88  6.48
y5     0  0.00  0.00   0.00 138.38  0.00  0.00
y6     0  0.00  0.00   0.00   0.00  0.00  0.00
y7     0  0.00  0.00   0.00   0.00  0.00  0.00
```

An F-value=2.85, having 1 and around 90 degrees of freedom (length of time series), corresponds to a p-value$\simeq$10% . Therefore, the dependence of the legislation is only highly significant for variables 2 (suicides with firearms) and 4 (homicide with firearms) with p-values below 1% (1-pf(6.48,1,90)$\simeq$1.3%).

Further, it is seen that variable 5 (killings without using firearms) does not seem to depend on any of the other variables (2, 3, 4), neither in the autoregressive nor in the moving average part of the model:

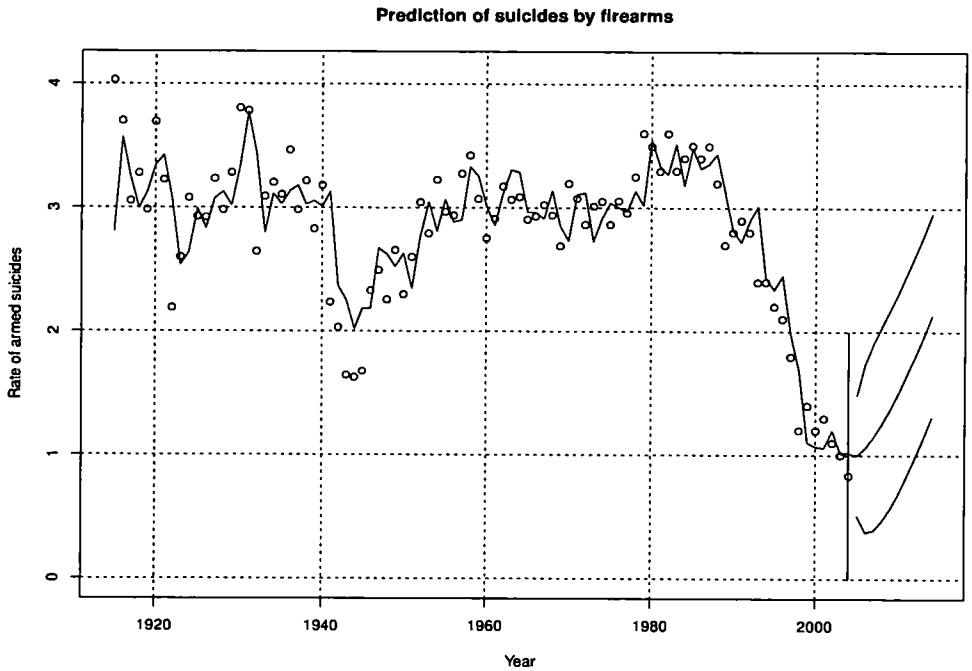
$$(y_{5,t} - 1.104) - 0.5619 \cdot (y_{5,t-1} - 1.104) = u_{5,t} - 0.7163 \cdot u_{5,t-1}$$

in that the mean of the observed y_5, 1.104, was subtracted from the observations before the marima estimation.

7.4 Prediction of timeseries

The routine called `arma.forecast` is used. We start by estimating our model (as before), and then we use the routine `arma.forecast`.

The data (o), the 1-step-ahead forecasts (–) and the nstep=10 forecast (–) and a 90% prediction interval for the forecast are shown in the plot below. Note, that the prediction interval is computed from the marima estimates and without taking the estimation uncertainty into account.



7.5 Forecasting variance

If a forecast $\widehat{y_{t+\ell}}$ over ℓ time units is calculated ($\ell \geq 1$), the variance of that forecast will be $\mathrm{Var}\{\widehat{y_{t+\ell}}\} = \sum_{i=0}^{\ell-1} \psi_i \Sigma_u \psi_i^T$, which can be derived from equation 7. The prediction interval shown in the above plot is calculated using this equation.

---

8 References

[1] Baker, J. & McPhedran, S. (2007) Gun Laws and Sudden Death, British Journal of Criminology, 47: 455-469.

[2] Jenkins, G.M. & Alavi, A.S. (1981) Some Aspects of Modelling and Forecasting Multivariate Time Series, Journal of Time Series Analysis, Vol. 2, no 1.

[3] Madsen, H. (2008) Time Series Analysis, Chapmann & Hall (in particular chapter 9: Multivariate time series).

[4] Reinsel G.C. (2003) Elements of Multivariate Time Series Analysis, Springer Verlag, 2^{nd} ed. pp 106-114.

[5] Spliid, H. (1983) A Fast Estimation Method for the Vector Autoregressive Moving Average Model with Exogeneous Variables, Journal of the American Statistical Association, Vol.78, no.384.

Parkometre i Fælledparken og FCK's hjemmekampe

Anders Milhøj
Økonomisk Institut,
Københavns Universitet
anders.milhøj@econ.ku.dk

Resume: I artiklen studeres datasæt med alle transaktioner på to udvalgte parkometre i Fælledparkområdet i København. Et af de udvalgte parkometre er beliggende lige over for Telia Parken på Øster Allé 33, hvor der som regel er enkelte biler parkeret. Det andet parkometer er beliggende på Edel Sauntes Allé, hvor der normalt ikke parkeres meget. Hele Fælledparken er dog belastet af mange parkerede biler, når der arrangementer i Telia Parken, fx koncerter men i sagens natur især ved FCK's hjemmekampe

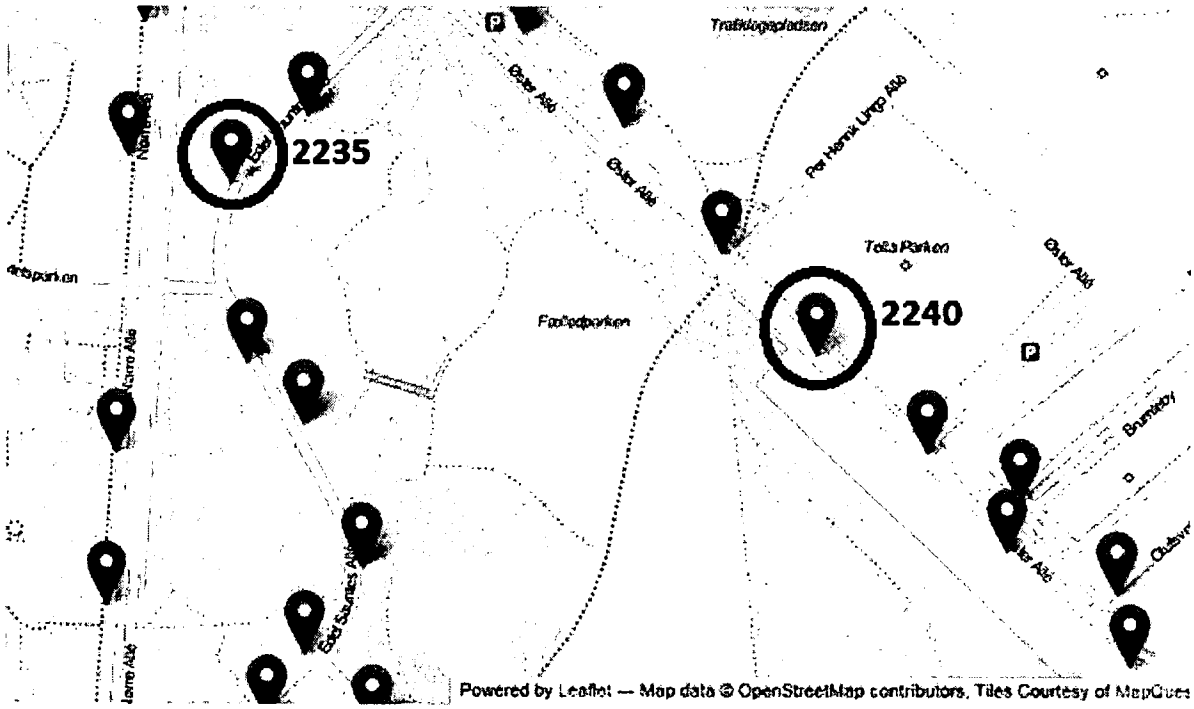
Indledning

Københavns Kommune stiller en lang række datasæt til rådighed for offentligheden på hjemmesiden data.ku.dk. Et eksempel er datasæt med alle transaktioner på alle parkometre i København i årene 2010 - 2014. Fra dette datasæt anvendes kun observationer fra to parkometre i Fælledparken. Datasættene for de to parkometre er i sig selv overvældende med visse fejl. Transaktionerne aggregeres til time-niveau og til dag-niveau hvorefter de analyseres med en SAS procedure for Unobserved Components Models. tidsrækker for parkometre er jo påvirket af almindelig parkeringsadfærd, fx er det gratis at parkere på søndage og visse helligdage.

Mange sære begivenheder har påvirket parkeringsadfærden i løbet af den betragtede tidsperiode, fx skal jo kun et mindre vejarbejde til for at udelukke parkering i en periode. Den slags er det i praksis umuligt at spore for en historisk periode. Andre begivenheder fx at hele Fælledparken spærres af ved DHL stafetten i august og også den omfattende parkering, der følger med arrangementer i Telia Parken, især FCK's hjemmekampe, let kan identificeres.

Fælledparken

Figuren viser et kort over dele af Fælledparken med markering af alle parkometre. De to parkometre, der anvendes i denne fremstilling, er markeret på kortet.



De to udvalgte parkometre er dels et parkometer lige over for Telia Parken ud for Øster Allé 33. I bygningen på Øster Allé er der et advokatkontor og andre liberale virksomheder, der giver anledning til parkering i dagtimerne. Desuden er FCK's fanshop beliggende lige ved parkometre, så kunder til den butik også med fordel kan anvende netop dette parkometer. derimod er der et stykke vej til områder med mange privatboliger, så der er kun undtagelsesvist parkering i forbindelse med privatbesøg i folks hjem. Som regel er det let at finde en parkeringsplads der. Ikke desto mindre er der i datasættet foretage ca 35.000 transaktioner på parkometeret i løbet af de fire år 2010-13.

Det andet parkometer ligger på en øde strækning af Edel Sauntes Allé. Normalt den slags gøremål. Gæster i Fælledparken fx motionister kan parkere der, men det er normalt let at finde en parkeringsplads der. Der er kun registret ca 5000 transaktioner idet perioden dog kun er på 1½ år, fra 1/6 2012 til 31/12 2013.

Aggregering

Data indeholder observationer for alle transaktioner, dvs oplysning om tidspunktet for transaktionen of oplysning om hvor lang tids parkering, der er betalt for. Denne tid kan være meget lang, da en times parkering købt lørdag eftermiddag kl 16.01 er gyldig frem til mandag morgen kl 8.01, da parkering er gratis på søndage. Ofte betales der

derfor for parkeringstid, der ikke udnyttes. Det er derfor valgt udelukkende at se på antal transaktioner.

Antal transaktioner aggregeres ved PROC TIMESERIES i SAS til hhv time-niveau og til dag-niveau. I perioder hvor der ikke har været en eneste transaktion er der naturligvis nul transaktioner. Tidsrækkerne på dag-niveau og især på time-niveau indeholder derfor en lang række nuller. Et højt antal transaktioner kan opstå hvis mange parkerer i meget kort tid, så der er også enkelte ret store observationer.

Antallet af transaktioner inden for et tidsinterval kan (til nød) opfattes som en observation fra en Poisson fordeling; dog stærkt zero-inflated. Som en transformation, der kan stabilisere variansen og som kan hjælpe lidt i retning af at få opfyldt en normalfordelings antagelse, anvendes derfor en kvadratrods transformation.

Unobserved Component Models

I denne fremstilling anvendes Unobserved Component Models i form af SAS-proceduren PROC UCM, se Milhøj (2013). Disse modeller er som udgangspunkt State Space Modeller, hvori parametrene estimeres ved Kalman filteret, men SAS-proceduren kræver ikke kendskab til de nærmere detaljer i modelleringsteknikken.

I modellerne lægges forskellige delkomponenter sammen til en samlet model. I dette tilfælde anvendes først og fremmest en niveau-komponent. Denne komponent er dynamisk, dvs at den principielt er af formen

$$\mu_t = \mu_{t-1} + \eta_t$$

hvor μ_t er selve komponenten og η_t er et restled. Restleddet opfattes om uafhængige normalfordelte variable med middel nul og en positiv varians. Jo større denne varians er, jo mere kan niveauet variere. Da der anvendes højfrekvente observationer, dvs daglige eller endog timelige observationer, er en eventuel årsvariation i parkeringsmønsteret umuligt at modellere parametrisk. I stedet vil årstidseffekter i parkeringen give sig udslag i ændringer af denne niveaukomponent μ_t. Variansen på restleddet estimeres ud fra data.

Ud fra denne niveau komponent defineres så selve tidsrækken x_t ved

$$x_t = \mu_t + \varepsilon_t$$

hvor igen ε_t er en følge af uafhængige normalfordelte variable med middel nul og en varians, der estimeres ud fra data.

For tidsrækken af transaktioner aggregeret til timeniveau er det oplagt, at der forekommer autokorrelationer. Er der købt mange parkeringsbilletter en time, er de ringen (eller højst få) ledige parkeringspladser tilbage, og derfor vil der blive købt færre billetter timen efter. Denne negative autokorrelation kan modelleres ved at

opfatte restleddet ε_t som en ARMA model. I analyserne i dette papir anvendes en AR(3) model for restleddet.

Der er naturligvis en kraftig sæsoneffekt i parkeringen målt på dagligt niveau, især da parkering er gratis på søndage, så transaktioner på søndage sjældent forekommer. Derfor tilføjes sæsondummy variable for dageffekter. Det er muligt at lade sæsoneffekter variere over tid i den anvendte modelklasse, men sæsonstrukturen viser sig i data at være konstant, så denne facilitet udnyttes ikke.

Desuden parkeres der yderst sjældent om natten, så for de timeligt aggregerede observationer skal der udover en ugedagseffekt også tillægges timeeffekter for døgnets 24 timer.

I selve grundmodellen anvendes ikke yderligere komponenter eller forklarende variable. Den anvendte model er derfor i sig selv utilstrækkelig, da en stor del af variationen i den observerede tidsrække x_t data skyldes kendte effekter, der ikke uden videre kan opfattes som dele af de stokastiske restled ε_t og η_t.

For dette datasæt er der nemlig så mange deterministiske effekter til stede i data, at det er umuligt at tage højde for alle. Selvfølgelig kan man godt tage højde for de få søgnehelligdage, hvor parkering er gratis, fx i forbindelse med Påsken. Men midlertidige afspærringer af Øster Allé i forbindelse med e motionsløb, der arrangeres af mange forskellige organisationer er en umulig opgave. På samme måde med midlertidigt parkeringsbesvær pga vejarbejde og andre fysiske hindringer for parkering - det er umuligt at få at vide om parkeringspladserne har været spærret en formiddag for et par år siden.

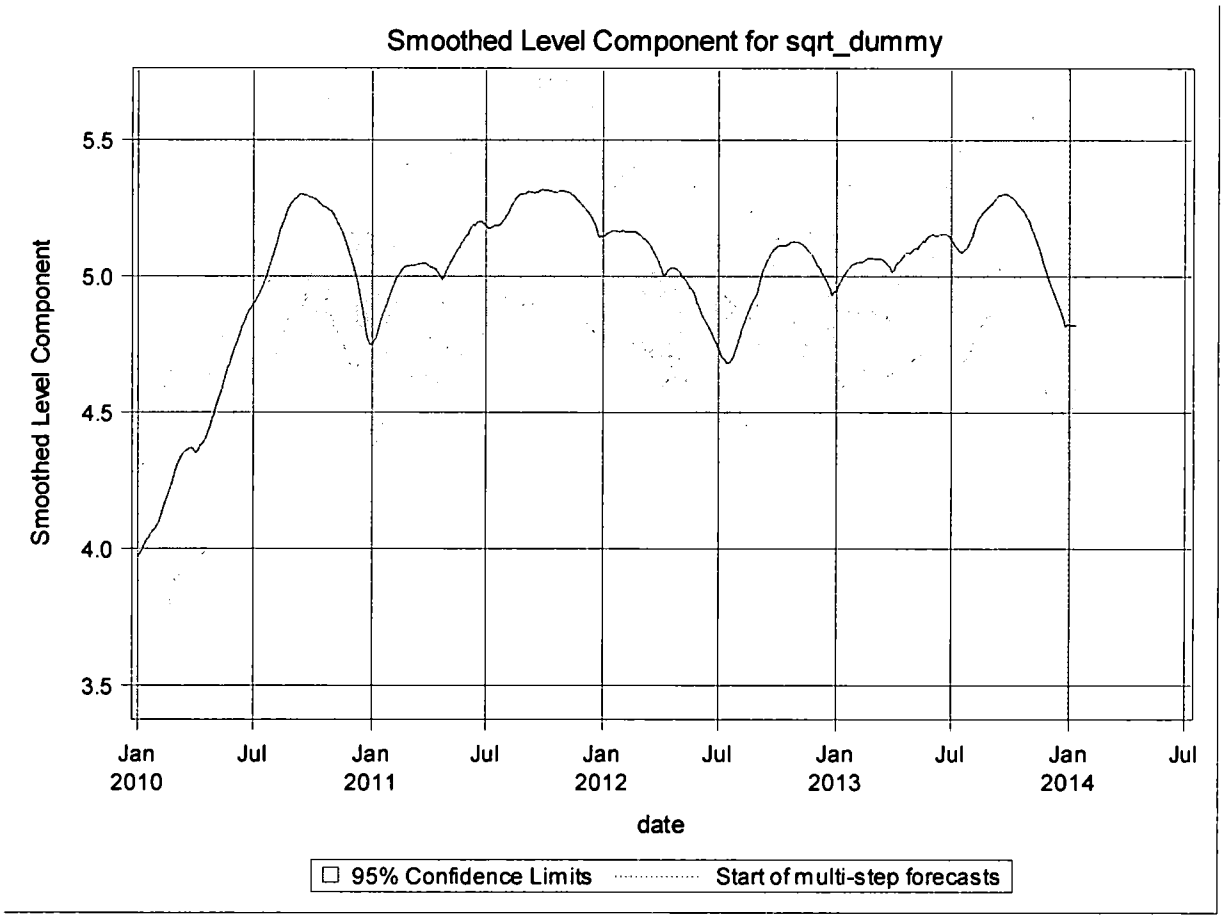
I modellerne skal man altså betragte de stokastiske restled som en skraldespand for alt det, der ikke er muligt at medtage i den deterministiske del af modellen. Dette er ofte et vilkår, når der skal analyseres store mængder af ikke-eksperimentelle data. Denne usikkerhed skal så tænkes ind i modelvalget - her vælges derfor en adaptiv model, hvor niveauet μ_t tillades at variere over tid. Yderligere undersøges alle større afvigelser fra modellen, dvs alle større restled ε_t. I det konkrete tilfælde sammenholdes større positive restled, dvs timer og dage med ekstraordinært mange parkeringer med viden om arrangementer, fx fodboldkampe, i Telia Parken.

Antal daglige transaktioner på Øster Allé

I dette afsnit betragtes antal transaktioner på parkometer nummer 2240 ved Øster Alle 33 aggregeret til dagligt niveau. Der anvendes data for alle dage fra og med 2/1 2010 til og med 31/12 2013. Antal transaktioner transformeres med en kvadratrod for at stabilisere variansen.

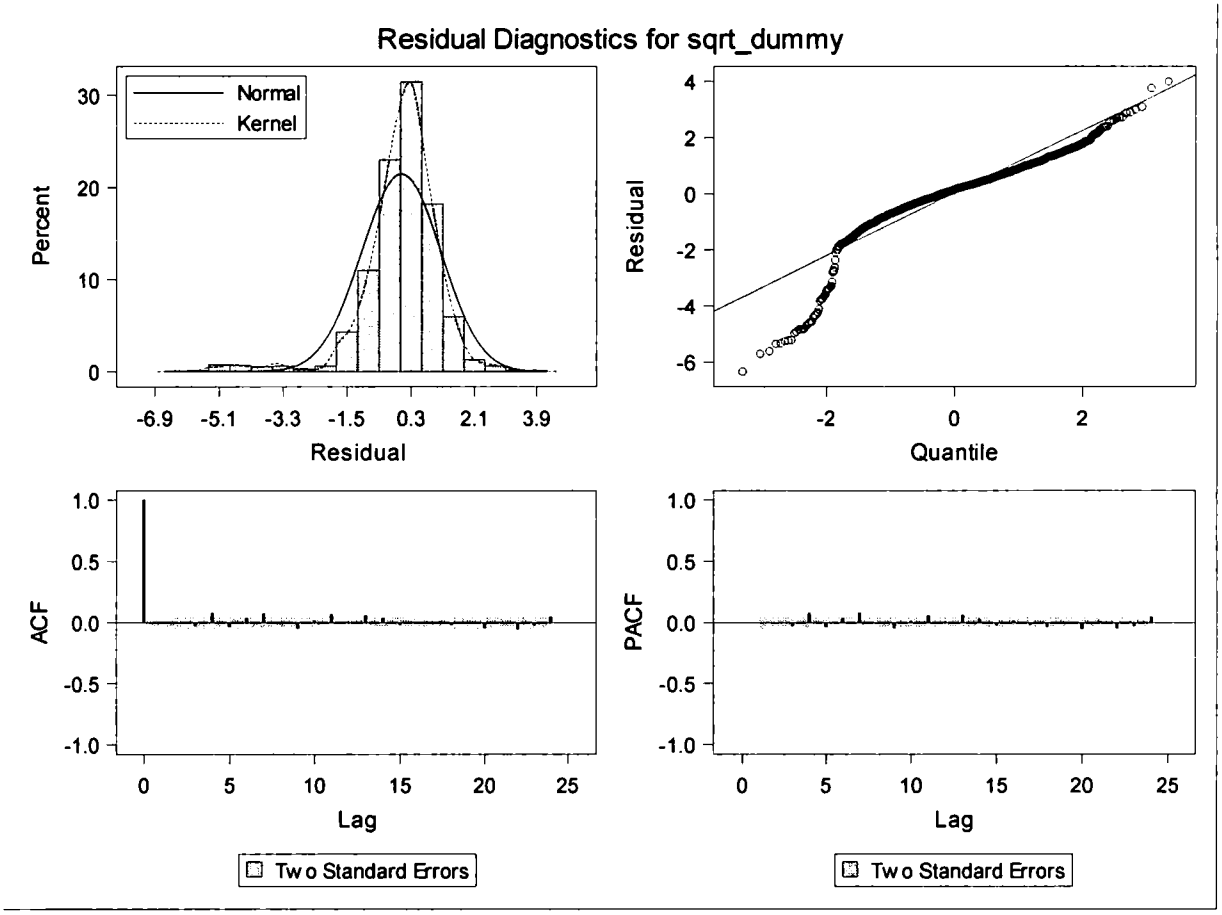
I modellen anvendes en tidskonstant ugedagsstruktur samt en tidsvarierende niveaukomponent samt en AR(3) model for restleddene for at afhjælpe visse mindre autokorrelationsproblemer.

Den tidsvarierende niveau komponent er som vist i figuren



Modellen har faktisk er udmærket fit som vist i figuren, der ikke viser problemer med autokorrelationer. Men der er en overvægt at små residualer, dvs dage med uventet få transaktioner, jf QQ-plottet.

Hvis der defineres en dummy variable for alle FCK's hjemmekampe kan den anvendes som højresidevariable i modellen. Den bliver faktisk signifikant med p =2.1% , med den forventede positive regressionskoefficent. Den forbedrer ikke den dårlige normalfordelingstilpasning.



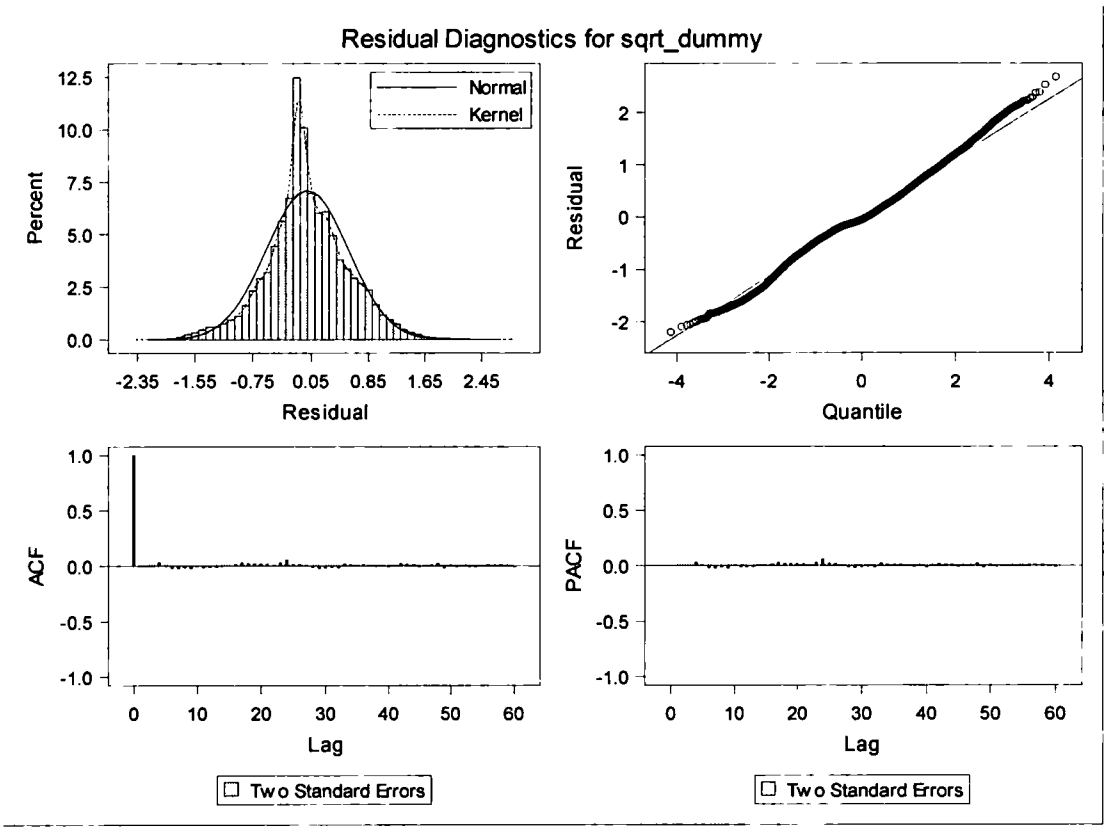
I programmet identificeres ene række outliere, hvoraf de mest signifikante er vist i tabellen nedenfor. Bemærk at der i tabellen kun er negativt afvigende observationer, dvs dage med uventet få transaktioner, men ingen dage med uventet mange transaktioner. Disse dage er alle fx Påske, Jul og den slags. Den eneste positive outlier i det viste udsnit er for 23/12 2012, hvilket vel sagtens kan forklares ved en del kortvarige julegave indkøb i FCK's fanshop.

date	*Estimate*	*Stddev*	*Chi-Square*	*Pr > ChiSq*	*Årsag*
13JUN2011	-5.65643	1.0344925	29.90	<.0001	Anden Pinsedag
01JAN2013	-5.57919	1.0344941	29.09	<.0001	Nytårsdag
01APR2013	-5.41624	1.0344924	27.41	<.0001	Anden Påskedag

Antal transaktioner på Øster Allé på timebasis

I dette afsnit betragtes antal transaktioner på parkometer nummer 2240 ved Øster Alle 33 aggregeret til time-niveau. Der anvendes data for alle dage fra og med 3/1 2010 til og med 31/12 2013. Antal transaktioner transformeres med en kvadratrod for at stabilisere variansen.

Grundmodellen giver et fint fit - ja selv normalfordelingstilpasningen er imponerende i betragtning af at der anvendes lidt over 35.000 observationer.



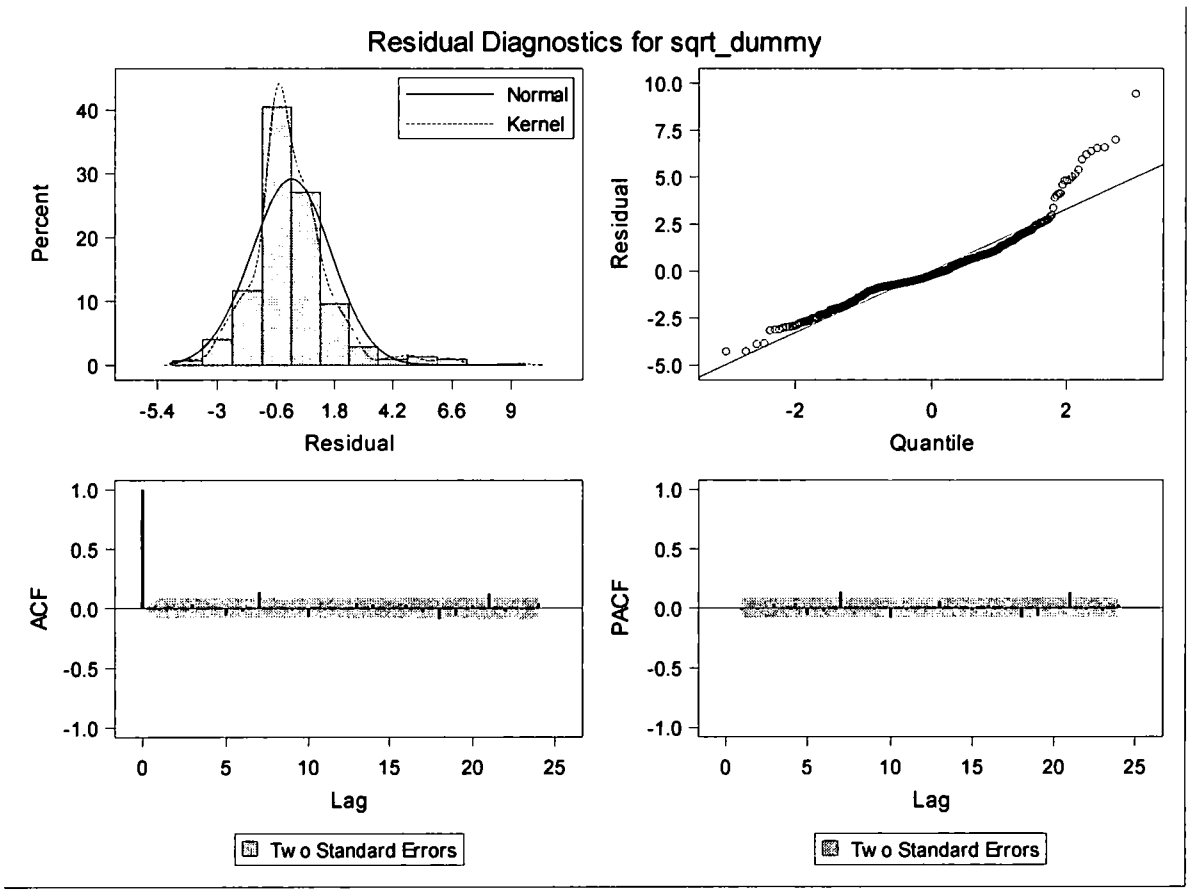
Den estimerede niveau komponent er nogenlunde som for de daglige observationer

Listen over outliere er lang, så der vises kun et udpluk

tlPayDateTime	*Estimate*	*Stddev*	*Chi-Square*	*Pr > ChiSq*	
24JUL10:14	2.74051	0.541544	25.61	<.0001	Pink
02AUG13:22	2.58370	0.5415709	22.76	<.0001	Night Glow (luftballoner) i Fælledparken kl. 22.00
16SEP10:15	2.33466	0.541544	18.59	<.0001	?
31DEC13:08	2.31045	0.5417569	18.19	<.0001	Nytårsaften?
18FEB12:12	2.30309	0.541544	18.09	<.0001	?

Antal daglige transaktioner på Edel Sauntes Allé

I dette afsnit betragtes antal daglige transaktioner på parkometer nummer 2235 på Edel Sauntes Alle aggregeret til time-niveau. Parkometret er placeret sted uden bebyggelse på en stille ensrettet vej gennem parken med mulighed for parallelparkering i begge sider af vejen. Der anvendes data for alle dage fra og med 1/2 2012 til og med 31/12 2013. Antal transaktioner transformeres med en kvadratrod for at stabilisere variansen.



Modellen giver et fint på nær at der en overvægt af positive store residualer. Disse residualer stammer fra diverse begivenheder i fælleparken, der har givet anledning til betalt parkering.

I modellen for disse antal daglige transaktioner indgår en regression med FCK's hjemmekampe som forklarende variabel. Denne komponent er selvfølgelig stærkt signifikant med en t-værdi på t = 9.89. Da denne komponent indgår i modellen er de positive residualer ikke foranlediget af FCK-hjemmekampe, men i stedet af andre begivenheder i Telia Parken eller i Fælledparken i øvrigt. Det fremgår tydeligt af tabellen over de mest signifikante begivenheder, der involverer koncerter, en landskamp og en europæsik klubkamp, de rike involverede FCK.

Date	*Estimate*	*Stddev*	*ChiSq*	*Pr > ChiSq*	*Årsag*
20APR2013	9.54799	1.516314	39.65	<.0001	Justin Bieber
26MAR2013	6.20102	1.5162644	16.73	<.0001	Landskamp - Bulgarien
20NOV2012	6.09200	1.5162968	16.14	<.0001	FC Nordsjælland spiller i Parken mod Shakhtar Donetsk
06JUN2013	6.05809	1.516273	15.96	<.0001	Bon Jovi

Antal transaktioner på Edel Sauntes på timebasis

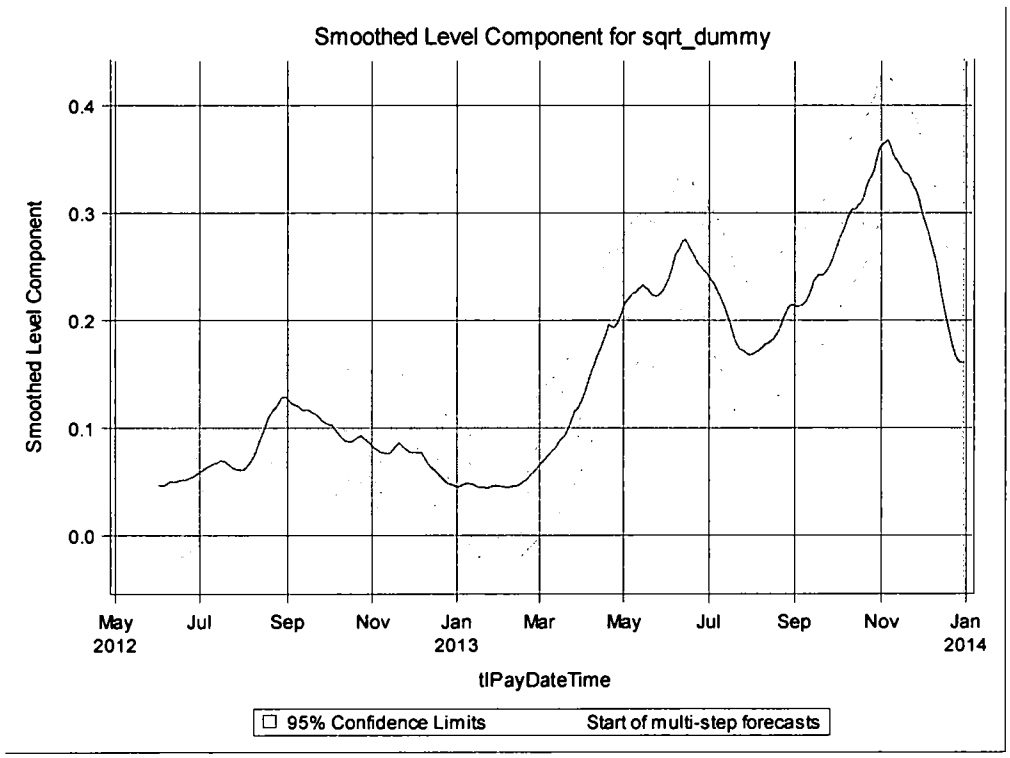
I dette afsnit betragtes antal transaktioner på parkometer nummer 2235 på Edel Sauntes Alle aggregeret til time-niveau. Der anvendes data for alle dage fra og med 1/6 2012 til og med 31/12 2013. Antal transaktioner transformeres med en kvadratrod for at stabilisere variansen.

Den estimerede model er en grundmodel, der kun tager højde for sæson på både ugedags- og timeniveau. Der er derfor mange outliere som vist i tabellen, idet stort set enhver anledning til parkering på Edel Sauntes Allé vil give anledning til så mange transaktioner, at det slår ud som en outlier. I modellen er der ikke taget højde for FCK's hjemmekampe i Telia Parken, så de fylder alle 7 mest signifikante på listen.

Fx er den mest signifikante outlier er 7/12 - 2013 i timen fra kl 16 - 17. Den dag spillede FCK efterårssæsonens sidste hjemmekamp mod FC Vestsjælland med 11179 tilskuere. Da december måned 2013 var i en periode med stort set ingen parkering i Edel Sauntes Allé jf plottet af level komponenten, gav den ekstra parkering anledning til en outlier.

tlPayDateTime	*Estimate*	*Stddev*	*Chi-Square*	*Pr > ChiSq*	*Årsag*
7DEC13:16	5.07483	0.3814373	177.01	<.0001	Vestsjælland
30OCT13:20	4.92976	0.3814354	167.04	<.0001	OB
06DEC12:18	4.12860	0.3814353	117.16	<.0001	FC Steaua Bucuresti
02NOV13:14	4.06937	0.3814352	113.82	<.0001	FC Nordsjælland
21AUG12:20	3.79236	0.3814353	98.85	<.0001	LOSC Lille
17SEP13:18	3.38408	0.3814353	78.71	<.0001	Juventus
15OCT13:18	3.37685	0.3814353	78.38	<.0001	Landskamp - Malta

Niveauet for level-komponenten er en anelse stigende i perioden, men den afspejler selvfølgelig, at der ikke parkeres meget på Edel Sauntes Allé.



Referencer

Anders Milhøj, 2013 Practical Time Series Analysis using SAS, *SAS Press*

Is there a fertility paradox in Denmark?

Jørgen T. Lauridsen, COHERE, University of Southern Denmark, jtl@sam.sdu.dk.

Summary

The present study analyzes the geographic variation across Danish municipalities in the fertility rate during the years 1982 to 2004. Several factors commonly believed to explain the variation in the fertility rate is found to be exerted to considerable regional variation. A model linking the fertility rate to several economic determinants is established and further modified to capture geographic small-area variation. A positive correlation between regional levels of income and fertility is found, which contradicts the so called fertility paradox.

1. Introduction

During the 20th century, the fertility rates in Western European countries have shown varying patterns during. Generally, the tendency is declining as concluded from the Princeton European Fertility Project by Coale and Watkins (1986). However, the general tendency covers considerable national and regional variation. This is demonstrated for the case of Denmark by Figure 1, which shows the development in the average of the fertility rates for 270 Danish municipalities from 1982 to 2004, including bands defined by plus/minus two standard deviations. Apparently, the average municipal fertility rate increases until the mid 1990'es, followed by a decline until 2000, where after it increases again. On the other hand, taking the bands in consideration, the rate is close to stable during the period.

Figure 2 shows the distribution of fertility across municipalities in 1982, 1993 and 2004. Throughout, fertility appears to move from peripheral and countryside to central and city areas.

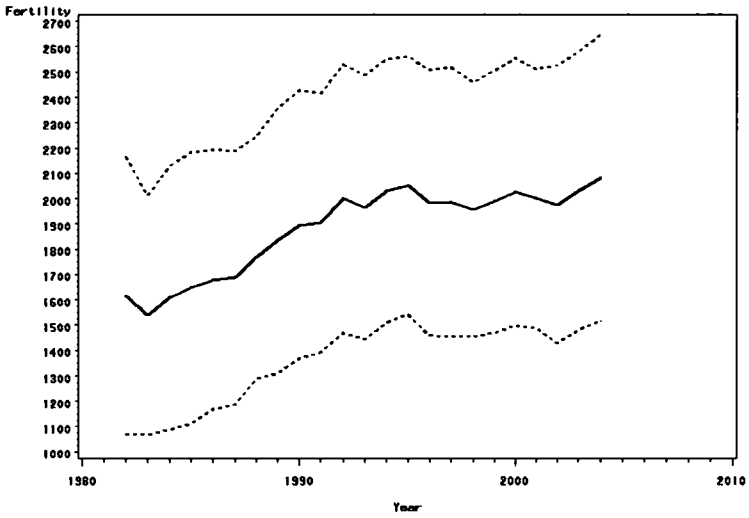


Figure 1. Average municipal fertility rates with +/- 2 std. bands.

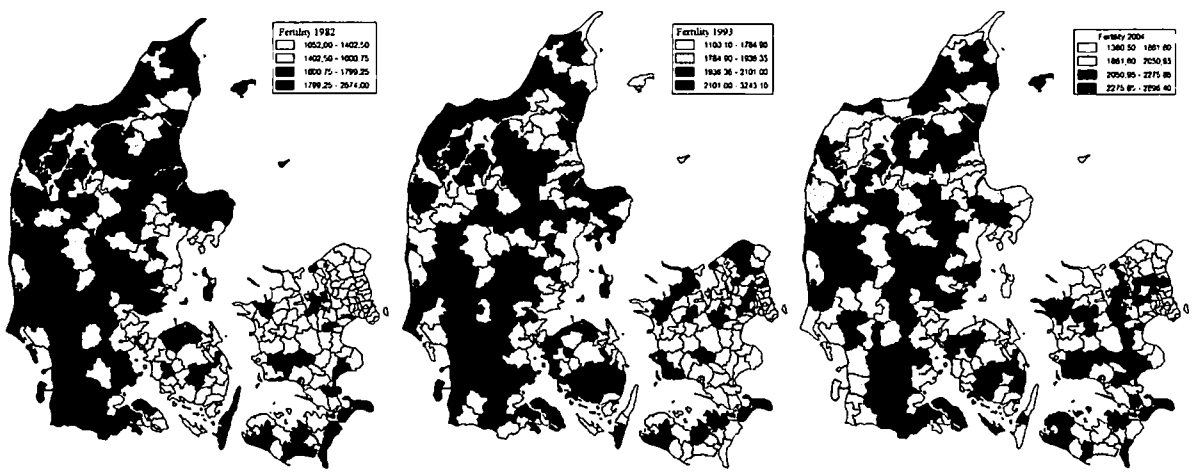


Figure 2. Fertility rates by municipalities

The purpose of the present study is to establish a model to explain the variation in fertility rates by variation in economic determinants. In particular, focus will be on the fertility paradox, i.e., whether there is a positive or negative relationship between income and fertility. Section 2 provides a review of economic theory, while Section 3 presents the methodology applied, including methods to adjust for geographical variation and spill over. Upon presentation of the data in Section 4, results of the

analysis are presented in Section 5, followed by concluding remarks to round off in Section 6.

2. Determinants and geography of fertility

From an economic perspective, households are assumed to be rational units acting optimally in any given situation in order to maximize the utility of the household. Children are assumed to provide a utility to the household which is compared to that of other material and non-material goods, including education, professional career etc. Thus, economic analyses predominantly focused on the number of children in the household (Ermisch 1991; Tasirin 1993; Hotz et al. 1997) or the timing and spacing of children during the woman's life cycle (Heckman and Willis 1976; Wolpin 1984; Moffit 1984; Cigno and Ermisch 1989; Tasiran 1993; James 1996).

Early studies (Leibenstein 1957; Becker 1960, 1965; Mincher 1963; Wilkinson 1973) stressed that the female wage rate is negatively correlated to the household's number of children, while the male wage rate is positively related. The argument is that the male is the primary income earner of the family, while the female has the main responsibility for the children. Therefore, the higher the female income, the more expensive it is to stay out of work. Empirical evidence is provided by several authors (Winegarden 1984; Lee and Gan 1989; Wang and Famoye 1997).

Given that the negative relationship between income and fertility to some extent conflicts with economic theory and intuition, it has been denoted the fertility paradox or the demographic-economic paradox (McFalls 1987; Weil 2004).

However, the support for a negative relationship has become weaker in most countries over recent decades and even turned into an expectation of a positive relationship (Siegel 2012). Denmark, alike other Nordic countries, is one example where the fertility and the female labour participation rate increased simultaneously until the early 1990'es. One reason is the build up of public welfare systems, including children day-care and financing of maternal leave. Moreover, the norm of most young people – male as well as female – has long been education first; next ensure position on labour

market, and finally having children (Sundström and Stafford 1992; Hoem 1993; Hoem and Hoem 1996; Andersson 1999).

On the other hand, while good economic conditions may be expected to correlate positively to fertility, the time spent for education delays the time until birth of first children and thus reduces fertility (Gustafsson et al. 2001; Löfström and Westerberg 2002; Cigno and Ermish 1989). It is therefore expected that the fertility rate will be higher in municipalities with a high rate of uneducated people.

Taking further the arguments that income may potentially increase fertility, it seems reasonable that unemployment reduces fertility.

Demography and family formation is another important set of factors behind fertility. It is expected that municipalities with many divorced individuals possesses lower fertility rates, and, correspondingly, that municipalities with many married people has high fertility rates (Alesina and Giuliano 2007). Ethnic minorities are known to establish more traditional family patterns, where the female takes care of the children and where the number of children is higher. Thus, a positive relationship between the number of persons from ethnic minorities and the fertility rate is expected (Lievesley 2010).

Geography and surroundings are known to be considerably connected with the variation of fertility rates, and spatial panel data methods have been applied (Galloway, Hammel and Lee 1994; Brown and Guinnane 2002; Guinnane 2011; Goldstein and Klüsener 2014). Housing units are more expensive in cities than in rural areas, so that families with children are motivated to move away from city areas, i.e., the degree of urbanisation is expected to be negatively related to fertility (Martine et al. 2013). Further, the service offered by the municipality, relative to the payment for this service via taxes, is an important factor which attracts families with children and thus increases fertility (Löfström 2000).

A final important matter is related to the small-area geographical dispersion of fertility. Specifically, if the number of families with children is large in a municipality,

then – everything else kept equal – a spatial spill over of families with children to neighbourhood municipalities may be expected. In other words, the fertility in a municipality is expected to be positively related to the average fertility in the municipalities surrounding it (Sandberg and Westerberg 2005). A look at Figure 2 illustrates that such a spatial spill over is present. Such a spatial spillover will be denoted an endogenous spillover, in order to distinct it from exogenous spatial spill over, caused by the explanatory factors (Anselin 1988). The latter notion refers to the phenomenon that the impact of these factors may go across the municipal border lines. For example, if a municipality offers a good service to child families, then the fertility will be relatively high. But what if the average level of service in the neighbourhood municipalities is high? Then these municipalities will be more attractive for families with children than the municipality considered, so that – everything else kept equal – the fertility will be lower in the municipality considered. Such exogenous spatial spill over effects may be of a contra-signed nature as illustrated by the example, but they may also be of the same sign as their non-spatial counterparts. As an example, income increases fertility. Likewise, high average income in neighbourhood municipalities will spill over to the municipality in the form of family movements, whereby fertility is increased.

3. Methodology

The point of departure is a linear regression model defined for the N=270 municipalities in a single year by

$$(1) \qquad y_t = X_t\beta + \upsilon_t, \quad \upsilon_t \sim N(0,\sigma^2 I)$$

where X_t is an N by K dimensional matrix of the K explanatory variables, y_t an N dimensional vector of the fertility rates in the municipalities, and β a K dimensional coefficient vector measuring the effects of the explanatory variables on fertility. The term υ_t is a residual term, which represents the fertility rates when controlled for the explanatory factors of X_t and may be denoted the residual fertility.

As discussed above, spatial spill over patterns across the municipalities have to be taken into account. Operationally, endogenous spatial spillover is controlled for by adding the average of y_t in the neighbourhood municipalities (denoted by y_t^W) as an explanatory variable in (1) to obtain the *spatially autoregressive* (SAR) specification (Anselin, 1988)

$$(2) \qquad y_t = y_t^W \lambda + X_t \beta + \upsilon_t,$$

where λ is a parameter specifying the magnitude of spill-over, formally restricted to the interval between (-1) and (+1), but for most practical purposes restricted to be positive. Likewise, exogenous spatial spillover is controlled for by adding the averages of X_t in the neighbourhood municipalities (denoted X_t^W) as explanatory variables in (1) to obtained the *spatially distributed lag* (SDL) specification (Florax, 1992)

$$(3) \qquad y_t = X_t \beta + X_t^W \delta + \upsilon_t,$$

while both types of spillover are controlled for simultaneously by simply involving both y_t^W and X_t^W to obtain a combined SAR-SDL specification.

Finally, between any two years, the covariance of the residual fertility reads as

$$(4) \qquad E(\upsilon_t' \upsilon_s) = \sigma_{ts}^2 \quad t,s = 1,..,T.$$

To obtain efficient estimates of β, we apply Feasible Generalised Least Squares (F-GLS) estimation as suggested by Zellner (1962) to obtain Seemingly Unrelated Regression (SUR) estimates for β.

4. Data

The data to be applied are defined in Table 1, which further shows the means by year of the variables. Data were obtained for 270 Danish municipalities annually from 1994-2003.

Table 1. Data applied for the study

Variable	Definition
Fertility	Summaric fertility rate per 10,000 females[1]
No education	% population without further education[2]
Tax-Service	Ratio of municipal service to tax collected, annual country average = 100[2]
Urbanisation	% population living in urban area[2]
Unemployment	% population without employment[2]
Foreigners	Number of inhabitants from countries outside EU, North America and Canada per 1,000 inhabitants[2]
Married	% population who are married[2]
Divorced	% population who are divorced[2]
Tax base	Income deductible for municipal and county taxation per inhabitant[2]

Source: [1] Statistics Denmark (www.dst.dk) and [2] the Key Figure Base (www.im.dk)

5. Results

Table 2 shows the spatially unadjusted model and the model adjusted for spatial spill over, applying the SAR-SDL framework as described above. Apart from the explanatory variables, their spatial counterparts and the spatial lag of fertility rates, a time trend and the square of the time trend is added to capture the U shaped development of the fertility rates across years.

Inspection of the spatially unadjusted model generally confirms the initial expectations regarding effects of determinants. The U shaped development of the fertility rates across years is confirmed by the significantly positive second order term of the time trend. Percentage without education is positively related to fertility, i.e., a negative relationship between education and fertility is proved. The service-to-tax rate exerts a positive impact, confirming that municipalities offering a good service attract families with children and thus experience a higher fertility. It is, however, noticed that the effect is not significant. Municipalities with a high degree of urbanisation experience a lower fertility as expected. Unemployment exerts the expected negative impact on fertility. High proportions of foreigners lead to high fertility rates. Marriage is positively related to fertility, and divorce negatively, as expected. Finally, the fertility paradox is questioned, as income is positively related to fertility.

Table 2. Unadjusted and spatially adjusted models of fertility

	Unadjusted model	Spatially adjusted model	
		Direct effect	Effect of spatial lag
Constant	1428.12*** (226.74)	2466.51*** (380.07)	
Time trend	-55.92*** (9.06)	-49.94*** (10.94)	
Time trend squared	4.74*** (0.67)	4.24*** (0.76)	
No education	6.50*** (1.52)	8.24*** (1.63)	-4.73* (2.62)
Tax-Service	0.98 (1.12)	3.08*** (1.11)	-5.80*** (2.05)
Urbanisation	-3.52*** (0.58)	-3.53*** (0.57)	0.76 (1.00)
Unemployment	-11.38*** (4.19)	-4.46 (5.95)	-2.89 (7.72)
Foreigners	1.80*** (0.49)	1.20** (0.50)	-0.47 (0.86)
Married	17.35*** (2.74)	23.78*** (3.19)	-22.04*** (4.45)
Divorced	-49.95*** (5.49)	-26.71*** (8.56)	-37.29*** (10.80)
Tax base	1.81*** (0.57)	-0.21 (0.69)	2.21** (1.06)
Spatial lag of fertility			0.06*** (0.01)
LogL	-15783.73	-15748.22	
AIC	31701.46	31646.44	

By comparing the model the unadjusted model to the spatially adjusted, however, the importance of adjusting for spatial spill over becomes evident. For the latter, it is especially noticed that the positive effect of the service-to-tax rate turns out to be highly significant. Moreover, the spatial lag of the service-to-tax rate comes out with a negative effect. I.e., if the service level is high in a municipality, then the fertility rate will be increased. But if the average service level in the surrounding municipalities is high, then the fertility rate will be reduced. An alike contra-signed pattern is found for proportion of uneducated: The direct effect is positive as expected. But if the proportion of uneducated in the surrounding municipalities is high, then a negative effect on fertility occurs. If lack of education is a large-area phenomenon

characterising an entire region of municipalities, then this region will be an economically peripheral region, which is not attractive for families with children, i.e., the fertility rate will fall. An alike contra-signed tendency is present for percentage of married, but for other reasons: If this percentage is high in the surrounding municipalities, then these municipalities are potentially attractive for married couples, whereby – everything else kept equal – such couples will move away from the municipality considered so that a drop in fertility is caused. Opposed to married, it is seen that the direct as well as the spatial effect of percentage of divorced are negative.

An important observation regarding the effect of income is called for. It is seen that the direct effect of income is not significantly different from zero, while the positive effect of income is rather caused by the income level in the surrounding region of municipalities. Thus, the positive effect of income is super-regional, rather than restricted to the local municipality. Specifically, this illustrates that the negative relationship as suggested by the fertility paradox is merely a dynamic and transitional feature, while the small-area effect of income reflects traditional economic theory by being positive.

Moreover, it is noticed that the direct effects of percentage of foreigners and urbanisation are as expected, while - not especially surprising - the spatial effects for these are not significantly different from zero. Finally, a positive spill over from the average of fertility rates in surrounding municipalities is found as expected.

A few comments to the quantities for comparison of models remain. The LR test for the spatially unadjusted SUR versus a simple linear specification rejects the latter in favour of the former. Further, the LR test for the spatially adjusted model versus the spatially unadjusted strongly supports the spatially adjusted model. Put together, these quantities strongly support the necessity of controlling carefully for the spatial nature of the data as well as for the repetition of observations across time.

6. Conclusions

Alike most of the Western world, the Danish fertility rate declined throughout the 20th century simultaneous to economic growth. This development, which conflicts with economic intuition, has been denoted the fertility paradox, and several studies have been devoted to resolve it. The present study analyzes the geographic variation across Danish municipalities in the fertility rate during the years 1982 to 2004. Several factors commonly believed to explain the variation in the fertility rate is found to be exerted to considerable regional variation. A model linking the fertility rate to several economic determinants is established and further modified to capture geographic small-area variation. A positive small-area correlation between regional levels of income and fertility is found, which contradicts the fertility paradox. Thus, the necessity of separating small-area and dynamic variation, aiming at obtain a proper interpretation of the link between fertility and its determinants, is demonstrated.

References

Alesina A, Giuliano P. 2007. Divorce, fertility and the value of marriage. Working paper.

Andersson G. 1999. Trends in Childbearing and Nuptiality in Sweden: A Periodic Analysis. Demography Unit, Stockholm University.

Anselin L. 1988. Spatial Econometrics: Methods and Models. Kluwer Academics, Dordrecht

Becker GS. 1960. *An Economic Analysis of Fertility'*. National Bureau of Economic Research, pp 209-231.

Becker GS. 1965. A Theory of the Allocation of Time. *The Economic Journal*, vol 75, nr 299, s 493-517.

Brown, John C., and Timothy W. Guinnane. 2002. "Fertility Transition in a Rural, Catholic Population: Bavaria, 1880–1910." *Population Studies*, 56(1): 35–49.

Cigno A, Ermisch J. 1989. A Microeconomic Analysis of the Timing of Births", *European Economic Review*, vol. 33, pp. 737-760.

Coale, Ansley J., and Susan Cotts Watkins, eds. 1986. *The Decline of Fertility in Europe: The Revised Proceedings of a Conference on the Princeton European Fertility Project*. Princeton: Princeton University Press.

Ermisch JF. 1991. Lone Parenthood: An Economic Analysis. NIESR Occasional Papers XLIV, Cambridge: Cambridge University Press.

Florax RJGM. 1992. *The University: A Regional Booster? Economic Impacts of Academic Knowledge Infrastructure*. Aldershot: Avebury.

Galloway, Patrick R., Eugene A. Hammel, and Ronald D. Lee. 1994. "Fertility Decline in Prussia, 1875– 1910: A Pooled Cross-Section Time Series Analysis." *Population Studies*, 48(1): 135–58.

Goldstein JR, Klüsener S. 2014. Spatial Analysis of the Causes of Fertility Decline in Prussia. Population And Development Review 40(3): 497–525.

Guinnane T. 2011. The Historical Fertility Transition: A Guide for Economists. *Journal of Economic Literature* 49, 589–614.

Gustafsson S, Kenjoh E, Wetzels C. 2001. The Role of Education in Postponement of Maternity in Britain, Germany, the Netherlands and Sweden", *Working Paper*, Amsterdam University, Dep. of Economics.

Heckman JJ, Willis RJ. 1976. Estimation of a Stochastic Model of Reproduction: An Econometric Approach". *Household Production and Consumption*, Ed., Terleckyj, Nestor, E., Columbia University Press, National Bureau of Economic Research, New York, s 99-138.

Hoem JM. 1993. Public Policy as the Fuel of Fertility: Effects of a Policy Reform on the Pace of Childbearing in Sweden in the 1980s". *Acta Sociologica*, vol 36, s 19-31.

Hoem B, Hoem JM. 1996. Sweden's Family Policies and Roller-Coaster Fertility". *Working Paper*, Stockholm.

Hotz VJ, McElroy SW, Sanders SG. 1997. The impacts of teenage childbearing on the mothers and the consequences of those impacts for governments. In R. A. Maynard (Ed.), *Kids having kids: Economic costs and social consequences of teen pregnancy* (pp. 55-94). Washington, DC: The Urban Institute.

James HS Jr. 1996. The Impact of Female Employment on the Likelihood and Timing of Second and Higher Order Pregnancies". *Working Paper*, Department of Economics, University of Hartford, England.

Lee KCD, Gan CL. 1989. An Economic Analysis of Fertility, Market Participation and Marriage Behaviour in Recent Japan". *Applied Economics*, vol 21, s 59-68.

Leibenstein H. 1962. *Economic Backwardness and Economic Growth"*. John Wiley & Sons, Inc, USA.

Lievesley N. 2010. The future ageing of the ethnic minority population of England and Wales. London: Runnymede and the Centre for Policy on Ageing.

Löfström Å. 2000. Att sätta barn till världen – Födelsetal och arbetsmarknad, en förstudie", *Umeå Economic Studies*.

Löfström Å, Westerberg W. 2002. Economic Fluctuations and Fertility Swings in Sweden, 1965-1998, Umeå Economic Studies.

Martine G, Alves JE, Cavenaghi S. 2013. Urbanization and fertility decline: Cashing in on structural change. Working Paper, IIED, London.

McFalls JA Jr. 1987. Frustrated Fertility: A Population Paradox. New York; Oxford; Toronto and Melbourne: Oxford University Press.

Mincer J. 1963. Market Prices, Opportunity Costs, and Income Effects". *Measuremen in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehud Grunfeld*, ed. Carl Christ et al, Stanford, s 67-82.

Moffitt R. 1984. Profiles of Fertility, Labour Supply and Wages of Married Women: A Complete Life-Cycle Model". *Review of Economic Studies*, s 263-278.

Sandberg K, Westerberg T. 2005. Spatial Dependence and the Determinants of Child Births in Swedish Municipalities 1974-2002. Working paper.

Siegel S. 2012. Female Employment and Fertility - The Effects of Rising Female Wages. CEP Discussion Paper No 1156. London: CEP.

Sundström M, Stafford FP. 1992. Female Labour Force Participation, Fertility and Public Policy in Sweden". *European Journal of Population*, vol 8, s 199-215.

Tasiran AC. 1993. Wage and Income Effects on the Timing and Spacing of Birth in Sweden and the United States". *Ekonomiska Studier*, nr 35, Handelshögskolan vid Göteborgs Universitet.

Wang W, Famoye F. 1997. 'Modelling Household Fertility Decisions with a Generalized Poisson Regression, *"Journal of Population Economics"*, vol. 10, pp. 273-283.

Weil DN. (2004). *Economic Growth.* Boston: Addison-Wesley.

Wilkinson M. 1973. An Econometric Analysis of Fertility in Sweden 1870-1965". *Econometrica*, vol 41, s 633-642.

Winegarden CR. 1984. Women's Fertility, Market Work and Marital Status: A Test of the New Household Economics with International Data". *Economica*, vol 51, s 447-456.

Wolpin KI. 1984. An Estimable Dynamic Stichastic Model of Fertility and Child Mortality". *Journal of Political Economy*, vol 92, s 852-875.

Zellner A. 1962. An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias. *Journal of the American Statistical Association* 58: 977-992.

The Noise-to-Bias Illusion. Why a perfect model may looked biased when the noise level is high

Nicolai Skov Johnsen and Kaare Brandt Petersen
SAS Institute, Denmark

Symposium i Anvendt Statistik, January 2016

Abstract

In this paper we describe the phenomenon observed for some modelling tasks on interval target variables, namely that the amount of over- and underestimated data point is not balanced. For the low-target values there is an overweight of overestimated, while for the high-target values there is an overweight of underestimated. It looks like a modeling bias, but turns out to be mainly a consequence of the noise level and for that reason we have chosen to call it *The Noise-to-Bias Illusion.*

1 Introduction

Mathematical models for data analysis gets ever more popular as data gets more common. Many different fields of expertise such as economy, statistics, applied mathematics, and physics, has a history of using mathematical models and therefore many different names as evolved for the same concepts. In this paper we talk about *predictive models*, aka *supervised models*, that is, models $f(\mathbf{x};\theta)$ which by some example data set D and a cost function to be minimized $C(\theta;D)$, attempts to find the relation between some input vectors $\mathbf{x}$ and a corresponding output y,

$$\hat{y} = f(\mathbf{x};\theta), \qquad D = \{(\mathbf{x}_i, y_i)|i = 1, ..., N\}$$

When estimating a predictive model in supervised learning, no matter if it is linear regression, a decision tree, or a neural network, there are some diagnostic plots which is part of the evaluation process. Some of these diagnostic tools are model-specific as for example Cooks distance for estimating the influence of single data points in squared error linear regression [3] or saliency maps for a neural network [2].

Others diagnostic plots, however, are model-independent. These plots include Receiver-Operator-Characteristics (ROC), Lift curves and expected response for binary target variables and predicted vs target scatter plot or predicted vs residuals scatter plot for interval target models. Model-independent diagnostic plots are especially relevant when the focus of the task is not on the algorithmic aspects but the application. When focussing on the application, one single model – no matter how great it is – is not sufficient. Instead, multiple models must be tried out to compare the outcome of the different models in order to understand some of the key properties of the problem at hand.

The predicted vs residuals plot is often relevant because the cost function $C(\theta;D)$ to be minimized in order to find $\hat{\theta}$ given the data set D, very often includes assumptions about the

residuals, e.g. that they are normal distributed or homoscedastic. Thus, the predicted vs residuals scatter plot is an important plot in order to find out if assumptions of the model are fulfilled, and in that sense, to find out if we as professionals estimating models did our modelling work appropriately. As it turns out, however, there are other plots which are at least as important if you have a different role.

2 A problem of unbalanced residuals

Imagine that you are the owner of a company buying and selling houses. One way to estimate the right price could be to let real estate brookers inspect the house and send in a report and an estimate. However, assume you want your business to run without real estate brokers because of the expenses, the time of estimation or the subjective nature of the evaluation. Instead you want an algorithm to estimate the value of each house. So you hire an expert in mathematical models to work on the problem. But when shes finished - what kind of reports on the project would you need?

You would of course like to know the precision – when the model estimated 100.000 for a house can you then trust that number? Assume that you are informed that the uncertainty is 16.000 on average (mean absolute difference). You would typically also like to know how this precision depend on other things like zip-code, house size, lot size, real price, etc. The last part with the dependence on the price level is important, because of course it is relevant to know if the model is best for cheap house or expensive houses. By these considerations alone, you are as business owner moving beyond the standard diagnostics and the effects of that is quite interesting as we shall see.

2.1 Data

To illustrate the point, we have an example on a small data set on house prices. It is the Windsor data set by The Corporation of the City of Windsor, and consists of 546 sold houses along with some information about each. Compared to what in Denmark is know about each house through BBR[1] and OIS[2] the input variables are quite limited, but the data is merely serving as an example and the effect is the same on data using detailed information like the one from BBR and OIS. A short data profiling on the Windsor Housing data set can be found in Tabel 1.

2.2 Model

The model chosen is rather simple. Due to the heavy tailed nature of real estate price data, the target variable is log-transformed, $q_i = \log(y_i)$, providing the predictive model $\hat{q}_i = \sum_d \theta_d x_{di}$. The cost function is the quadratic error on the log-transformed variables $\sum_i (\hat{q}_i - q_i)^2$. This gives rise to a set of estimated parameters θ and subsequently a predictive model which estimates y for any input $\mathbf{x}$.

2.3 Evaluation

As a part of the evaluation, a binned relative performance measure is constructed: The estimation is called *OK* if $|\hat{q}_i - q_i|/q_i \leq 0.01$, *Overestimated* if $\hat{q}_i$ is larger than this definition and

[1] Bygnings- og boligregisteret – a Danish national database on building characteristics

[2] Den Offentlige Informationsserver – a Danish national database on real estate informations

Variable	Type	Role	Information
Sales price of house (USD)	Interval	Target	min=25000, max=190000, mean=68121
Size of lot	Interval	Input	min=1650, max=16200, mean=5150
Number of bedrooms	Interval	Input	min=1, max=6, mean=2.97
Number of bathrooms	Interval	Input	min=1, max=4, mean=1.21
Number of stories, excl basement	Interval	Input	min=1, max=4, mean=1.81
Number of garage places	Interval	Input	min=0, max=3, mean=0.69
Driveway	Binary	Input	86% no, 14% yes
Recreational room	Binary	Input	82% no, 18% yes
Full finished basement	Binary	Input	65% no, 35% yes
Gas for water heating	Binary	Input	95% no, 5% yes
Central aircondition	Binary	Input	68% no, 32% yes
Located in a preferred location	Binary	Input	77% no, 23% yes

Table 1: Variables on Housing data set. It is a rather small data set of 546 observations, but is illustrating the point of the paper nicely. The data has 11 input variables and there are no missing values for any of the input variables.

Interval (Target values)	# Obs	Underestimated	OK	Overestimated
87.000 - 190.000 USD	112	47 %	**46%**	6%
69.000 - 87.000 USD	104	35 %	**46%**	19%
57.000 - 69.000 USD	116	36 %	**45%**	19 %
47.000 - 57.000 USD	107	20 %	**50%**	30 %
25.000 - 47.000 USD	107	4 %	30%	**66 %**

Table 2: Evaluating the result by proportion of underestimated, well estimated, and overestimated in bins of the target output value. Note the clear relation between the intervals and the accuracy: The expensive houses are underestimated, while the lowprice houses are overestimated. (Largest ratio in bold))

Underestimated if $\hat{q}_i$ is smaller than this definition. This measure is standard within real estate valuation tasks and has the key property that it includes the entire price range because it is relative and that is insensitive to outliers. With this definition one can easily count the proportion of the data which falls into the three each of the three categories respectively. Doing this, the results are

	#Obs	Underestimated	OK	Overestimated
(All data)	546	28.7%	43.5%	27.8 %

Presented with this result it is natural to ask how this performance is on subsets of the data. As owner of a company buying and selling houses, there is one important question in particular: Is the model performing equally well on low-price estates as on high-price estates? In order to answer this question, the above proportion is reported on grouped parts of the target values and the result is seen in Table 2. The results are quite interesting. While the proportion of observations which falls within the OK-margin is reasonably constant over the target values, the proportion of Underestimated and Overestimated shows a clear pattern: For low-value targets, the model seem to overestimate far more often than underestimated (4% vs 66%); For high-value targets it is the other way around (47% vs 6%). Following the values in the columns

Interval (Predicted values)	# Obs	Underestimated	OK	Overestimated
87.000 - 190.000 USD	99	20 %	**53%**	27%
69.000 - 87.000 USD	102	32 %	**41%**	26%
57.000 - 69.000 USD	109	23 %	**49%**	28 %
47.000 - 57.000 USD	134	37 %	**37%**	27 %
25.000 - 47.000 USD	102	28 %	**41%**	30 %

Table 3: Evaluating the result by proportion of underestimated, well estimated, and overestimated in bins of the predicted output value. Note that there is no clear relation between the intervals and the accuracy. (Largest ratio in bold)

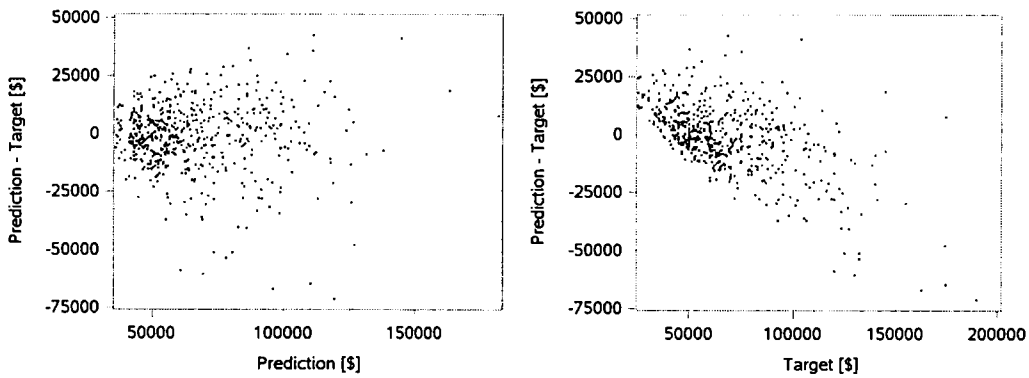


Figure 1: a) Predicted vs Residuals and b) Target vs Residuals. Note that while a) indicates no special structure apart form heteroscedasticity, plot b) shows a clear pattern of residuals being smaller as a function of target, from positive with small targets to negative for large target value.

of Table 2, there is no mistake possible: As the target values increase, the balance between Underestimated and Overestimated shifts from one extreme to the other.

It looks like a bias and it is natural to think that it is the result of a model misspecification. First check along those lines is a check on the same values but grouped by the predicted values in stead of the target values. The results are shown in Table 3 but the effect is no longer there; over the predicted values the under- and overestimation are reasonably balanced.

A different way to look into this is to scatter plot the residuals vs the target and the predicted values respectively and the results are seen in Figure 1. In these plots the effect manifests itself by the difference on the two plots: While the classical prediction vs residuals shows heteroscedasticity there is a balance between postive and negative residuals over the entire range of predicted values. For the target values, however, this is quite different with positive residuals to the left and negative residuals to the right. This reflects the same effect of overestimating the low-value targets and underestimated the high-value targets.

Summarizing, we have a classical estimation problem, which when subjected to a certain evaluation of the performance for different target values – an evaluation which is very reasonable from an application viewpoint – exhibits somewhat surprising properties.

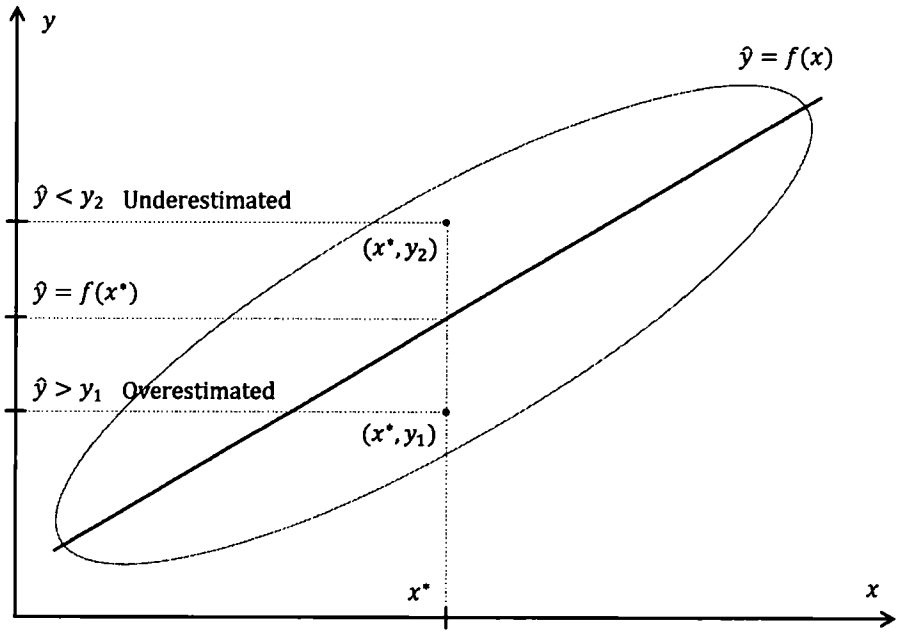


Figure 2: Conceptual view of data (grey cloud) with model (line). Note that for a specific input value on the x-axis, the model has the output value on the y-axis corresponding to the line. But the data cloud shows that for some data points (the points below the line), this will overestimate the value, while for other data points (the points above the line), this will underestimate the value.

3 Conceptual explanation

One way to better understand the effect is to look closer at the combined space of input and target. In a simplistic view, which lends it self well for illustrations, consider a space of one-dimensional input and a target variable, resulting in a two dimensional space. In this space, the data set D of pairs is a cloud of N data points.

Figure 2 illustrates this combined space of input and target: The first axis is the input variable, the second axis is the target variable and the data is illustrated as an elliptic cloud. The line is some estimated model, which for any new input value on the first axis, provides a model estimate by the corresponding value of the second axis. Note that for the data points in the data cloud with input value x^*, which has target values below $f(x^*)$ such as (x^*, y_1), the model will *overestimate* the target value y_1, since $y_1 < \hat{y} = f(x^*)$. Similarly, all data points like (x^*, y_2) with input value x^* but with target value y_2 larger than $f(x^*)$ will be underestimated. In general, the data points of the data cloud below the model line of $\hat{y} = f(x^*)$ will be overestimated and the data points above will be underestimated.

Figure 3 shows what happens when binning into intervals of the target variable and reporting on the performance. The binning with respect to target values naturally corresponds to horizontal slices of the data and in these horiosontal slices, the amount of data below and above the model line are not balanced: For the low-value parts of the target variable there is much larger area (more data) below the model line (colored red) than above (colored blue); and for the high-value target it is the other way around. Thus with a data cloud as the one sketched here, the unbalanced over- and underestimation is a natural consequence of the data cloud shape, the model and the way of measuring the performance.

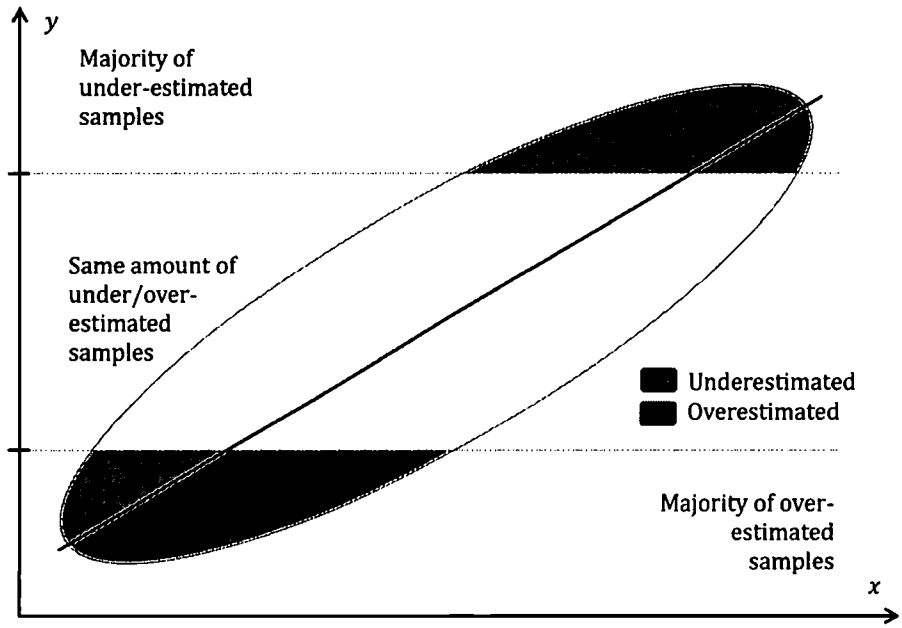


Figure 3: Conceptual explanation of unbalanced over- and underestimation. Note the size difference in the colored areas: For small target values, the red area (the overestimated) is much larger than the blue (the underestimated), and vice versa for the high target values. The larger area, the more data points it reflects, hence the unbalanced effect.

Figure 4 shows how the effect is related to the noise level. In the combined input/target-space, the noise level is in some sense the ambiguity of the target values for some values of the input-variables and in the 2d example, this corresponds to the 'thickness' of the data cloud. The figure shows that as the noise level increases the degree of unbalance gets more pronounced. The vice versa is also true: In the limit of no noise there is no longer an effect of unbalance.

The conceptual explanation above has a few subtle properties. Firstly the model drawn as a line through the data cloud is not as a standard linear model would actually fit. In the traditional linear regression model the error function is on the squared difference on the y-values only. The effect on the figures above is that the model line is less steep but this only enhances the described effect. The model line drawn is going through the top and bottom of the ellipse (the two points with strongest curvature), which corresponds to a so-called *orthogonal* estimation, minimizing the line distance from each data point to the model line – a distance which as a line is orthogonal to the model line.

Central to the conceptual explanation is the shape of the data cloud in the combined input/target-space. In the example above it is a classical elliptic shape which is a result of many processes and fits a two-dimensional Gaussian distribution. It is, of course, not the only possibility and since the effect of unbalanced performance depends strongly on the shape of the data cloud, one may have modelling challenges in which there is no unbalanced effect at all. In general, however, a data cloud leading to no effect at all is not very likely as it would require a certain trapezoid shape not seen very often in real life data sets.

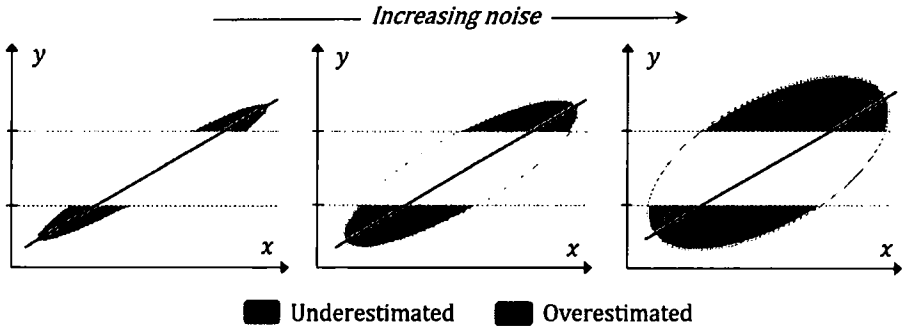


Figure 4: The influence of noise level. Note that as the noise level increases, and the data cloud gets thicker, the difference in size of the red and blue areas increase as well. The effect of unbalancing the underestimated and the overestimated is increasing with noise.

4 Investigation on synthetic data

While illustrations like in the previous section can be very beneficial for the understanding of the problem, they can also be misleading. One possible issue is misspecification of the model, which is highly likely bearing in mind the words of George Box that "all models are wrong, but some are useful"[1]. To investigate the effect of the Noise-to-bias illusion further, we have therefore constructed a setting in which we know that the model is not misspecified and we have done it using a so-called *generative function*, i.e. a known function which generates data.

The generative function and the construction of so-called synthetic data is

$$y_i = x_i + \sigma\epsilon_i, \qquad x_i \sim N(0,1), \epsilon_i \sim N(0,1)$$

where $x_i \sim N(0,1)$ means that x_i is a random variable from a normal distribution with mean value 0 and variance 1, and that it is independent over different values of index i. The parameter σ is controlling the signal-to-noise ratio. Constructing 5000 observations like this we get a data set of pairs $D = \{(x_i, y_i)|i = 1, ..., 5000\}$. When estimating the relation between x_i and y_i we know that in the limit of infinitely many observations, the estimated but correctly specified function must match the generative function above. Thus estimating $\hat{y} = ax$ using a quadratic error, which corresponds to the gaussian noise in a maximum likelihood approach, we have a model which is not misspecified.

Figure 5 shows the results for three different noise levels: $\sigma = 0.5, \sigma = 2$, and $\sigma = 5$. Both the generative model and the estimated are on the plots as lines, but since they are almost completely perfectly overlapping, they are indistinguishable for all three noise levels; this matches well with the setting to avoid misspecification. The top row shows input/target space in which the noise level shows itself as the thickness of the data cloud with clear differences for three chosen noise levels chosen. The bottom row shows the target vs the residuals, and the expected effect is very pronounced: As the noise increases the data cloud of residuals is 'rotating'. For high noise regime $\sigma = 5$ the residuals are positive for small target values and negative for large target values.

Thus, even in this simple and fully controlled setting with one-dimensional input, where the generative function is known and no misspecification ins influencing the results, the effect is still there. This leads us to conclude that al though alternative circumstances may also influence the results on real life data, the conceptual illustrations of section 3 are pointing out

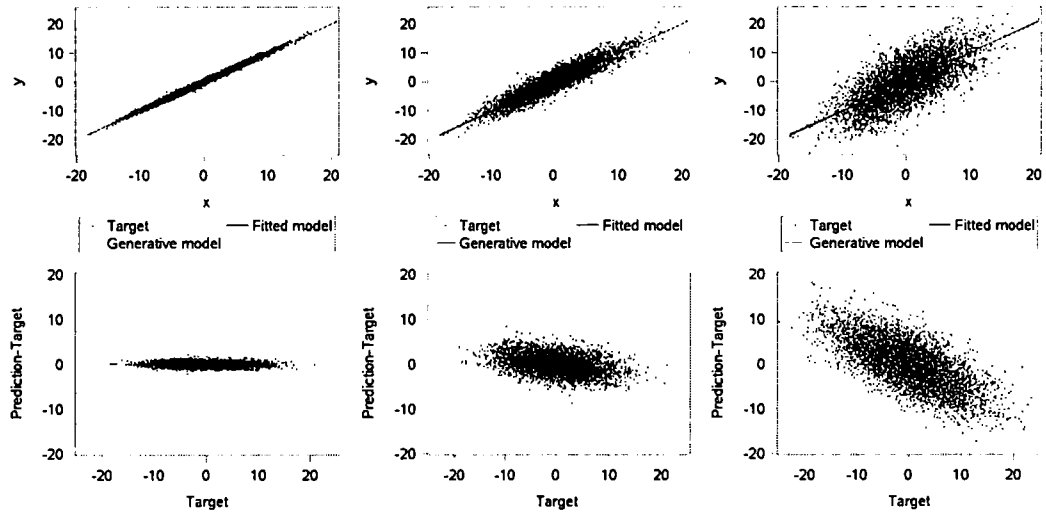


Figure 5: The influence of noise level on synthetic data. Note that as the noise level increases from left to right, the data cloud of the scatter plot in the target vs residuals, is rotating clockwise. This is the signature of a high noise level. In the upper row, there are two lines, al though they are almost unseparable visually: One is the generative model and the other the estimated. Thus, the estimated model has found the generative model (for all practical purposes), thus the effect is not due to mis specification of the model.

a valid observation: In high noise environments, the noise will cause the amount of over- and underestimated observations to be unbalanced.

5 Discussion

In summary, we have presented The Noise-to-Bias Illusion on a real life data set and demonstrated that is is not necessarily due to misspecification by reproducing the effect on synthetic data from a fully controlled setting with a correctly specified model. The effect is heavily dependent on the shape of the data cloud of the input/target space, and in so far that we have seen it in other real life modeling tasks, we conjecture that it is not uncommon. For that reason, we recommend as a standard diagnostic procedure to not only plot residuals versus the predicted values, but also the residuals versus the target values. That plot will tell if the noise level in the task at hand is so high that unbalanced estimation must be expected.

References

[1] Box, G. E. P. (1979), "Robustness in the strategy of scientific model building", in Launer, R. L.; Wilkinson, G. N., Robustness in Statistics, Academic Press, pp. 201236.

[2] "Neural Networks for Pattern Recognition", Christopher Bishop, Oxford University Press, 1995.

[3] Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," Technometrics, 19, 1518.

Minimal equivalent data sets: a prerequisite for a new platform for register-based epidemiological research

Klaus Rostgaard

klp@ssi.dk

ABSTRACT

A large fraction of epidemiological studies are entirely register-based and based on a very small repertoire of standard models. For reasons of safety and privacy the underlying data sets and the manipulations leading to the final model cannot be made public. Data sets for epidemiological research are often very big and very detailed, in large part due to the need for confounder adjustments. Hence the identity of some of the subjects under study would easily be revealed even without specific person identifiers to anyone with access to it. On the other hand for scientific reasons of transparency and documentation it would be desirable to be able to reproduce the inferences from the study based on a publicly available data set. This calls for the creation of "minimal equivalent data sets". These are data sets constructed in such a way that the form of the data is suited for analysis with the chosen standard model(s), and if so leads to exactly the same inferences regarding the interest parameters, while at the same time having only as many observations as there are unique combinations of levels of the predictors of interest. Here we explore the mathematical and practical feasibility of this approach.

INTRODUCTION

The writing of epidemiological papers is a large industry. Most papers conform to a very standardized format, with tight space limits and at the same time formally a desire for sufficient detail to allow reproduction and scrutiny of the results by peers. The statistical models used to transform data into inferences are almost always selected from a very small set of parametric or semi-parametric regression models. The industrial aspect of epidemiological research also calls for speedy, easy and cheap access to data etc. Denmark has one of the best legal and practical infrastructures for register-based epidemiological research in the world. Based on experience with many types of register-based research in Denmark, the following outlines a system that could arguably make the research and review process faster, cheaper, better documented and transparent, less error-prone, and take up less space in papers with technical details and formalia.

Anticipated system:

1. Raw data of person-level event histories (person, time, event, attributes)
2. ... are accessed, manipulated and analyzed through a menu-based interface on the web with a fixed set of standard options and survival analysis models at suitable institutions (e.g. Forskerservice at Statistics Denmark)
3. ... to produce output in the form of a "Minimal Equivalent Data Set" and documentation of the choices made in 2)
4. ... to be placed publicly available on the web for those who know the URL.
5. Legal and other requirements for the unfeasibility of identifying study subjects and populations are automatically checked and enforced by software during the creation of the minimal equivalent data sets in 2).

Here we focus on the creation of a minimal equivalent data set. This can be thought of as a statistically minimal sufficient data set for reporting an imaginary randomized trial analyzed with the same type of model as the actual study (e.g. Cox regression) but only involving outcomes, exposures of interest and relevant interactions, not all the other adjustments factors which randomization would have taken care of – and leading to the same statistical inferences. This condensation of data should by and large by itself secure the anonymity of the study participants.

HOW TO CONSTRUCT MINIMAL EQUIVALENT DATA SETS

We shall exemplify only with model(s) to be analyzed by so-called Poisson regression, where the underlying data are events, Y_i and time at risk ($PYRS_i$) for each combination of covariates x_i characterizing some follow-up time. We have the predictor vector $x_i=x_{1i},x_{2i}$ with corresponding parameter vectors θ_1, θ_2 where θ_1 is the interest parameter(s) and θ_2 represent nuisance parameters. Let the data set be $\{Y_i,x_{1i},x_{2i},O_i\}$ where $O_i=\log(PYRS_i)$ and sorted by x_{1i}. Define i* as the first i for which $x_{1i} = x_{1i*}$. Looking at the score equation reveals that the same maximum likelihood estimate is obtained for θ_1 when fitting the Poisson model on the data $\{Y_{i*},x_{1i*},O_{i*}\}$ where Y_{i*} is the summation over Y_i for which $x_{1i} = x_{1i*}$ and similarly $\exp(O_{i*})$ is the summation over the relevant $\exp(x_{2i}\theta_2+O_i)$ with θ_2 the maximum likelihood estimate for θ_2 from the big model. Similar reductions are possible for models based on logistic regression and Cox regression. A practical example is a Danish family study on infectious mononucleosis (IM) as a preliminary for a genome-wide association study, assessing whether your risk of IM, measured by hazard ratios is increased and by how much in case you have such-and-such type relative with a history of IM.[1] Here we adjust for sex-specific age, sex-specific calendar period, and birth cohort, all in one-year groups and also for having such-and-such type of relative. A minimal sufficient data set for assessing all the types of familial exposure considered contains N=390814 observations, while the minimal equivalent data set only contains N=51 observations.

Figure 1. Comparison of inferences from a minimally sufficient observational data set (Nobs=390814) for assessing familial risks for infectious mononucleosis and the corresponding minimal equivalent data set (Nobs=51) in a Poisson regression model with canonical link function (log).

A: Joint model

Obs	Parameter	Estimate	StdErr	equiv_ est	equiv_ StdErr	diffest	frac_std
1	Intercept	-9.3304	0.6697	-9.3304	0.0079	0.00000	0.01179
2	samesextwins7	2.2692	0.5851	2.2692	0.5774	-0.00000	0.98684
3	diffsextwins7	-15.9269	6588.696	-13.5844	2040.016	2.34246	0.30962
4	siblings7	1.1028	0.0778	1.1028	0.0773	0.00000	0.99330
5	parents7	0.6096	0.0730	0.6096	0.0726	0.00000	0.99507
6	halfsiblings7	0.6949	2.4405	0.6949	2.4402	-0.00000	0.99988
7	maternalhalfsiblings	0.2411	2.4406	0.2411	2.4402	0.00000	0.99985
8	paternalhalfsiblings	-0.5067	2.4406	-0.5067	2.4402	0.00000	0.99984
9	grandparents7	-0.5018	0.4474	-0.5018	0.4473	-0.00000	0.99982
10	uncles7	0.3282	0.0627	0.3282	0.0623	0.00000	0.99345
11	cousins7	-0.0015	0.0862	-0.0015	0.0855	-0.00000	0.99224
12	Scale	1.0000	0.0000	1.0000	0.0000	0.00000	.

B: Marginal models

Obs	Parameter	Estimate	StdErr	equiv_ est	equiv_ StdErr	diffest	frac_std
1	Intercept	-9.1150	0.6465	-9.3194	0.0078	-0.20443	0.01199
2	siblings7	1.1037	0.0778	1.1046	0.0773	0.00088	0.99335
3	Scale	1.0000	0.0000	1.0000	0.0000	0.00000	.
2	uncles7	0.3359	0.0627	0.3430	0.0623	0.00707	0.99312
2	samesextwins7	2.2316	0.5851	2.2412	0.5773	0.009576	0.98676

In this example the reduction in size of the relevant data set is dramatic. For all practical purposes the same inferences are obtained. And unless you happened to know a twin pair who had both been hospitalized for IM and one of these had a special feature in the data set (say an affected uncle) you would have learned absolutely nothing (in probability) from this data set about any identifiable person.

REFERENCES

1. Rostgaard, K., Wohlfahrt, J. & Hjalgrim, H. A genetic basis for infectious mononucleosis: evidence from a family study of hospitalized cases in Denmark. *Clin. Infect. Dis.* **58,** 1684–9 (2014).

Anders Milhøj
Nyheder i SAS

Department of Economics, University of Copenhagen
Øster Farimagsgade 5, DK-1353 København K
Anders.Milhoj@econ.ku.dk

I sommeren 2015 blev Analytical Produts opgraderet version 14.1, men stadig til Base SAS, version 9.4 sendt på markedet. Denne opdatering til nu Analytical Products 14.1 indeholder opdateringer af de analytiske programpakker indenfor statistik, økonometri, operationsanalyse etc. Disse opdateringer er nu løsrevet fra samtidige opdateringer af det samlede SAS-program. Desuden er den gratis SAS-applikation SAS-U opdateret; især er det bemærkelsesværdigt at der via SAS-U nu også stilles procedurer til rådighed for avancerede ikke-modelbaserede tidsrækkeanalyser.

SAS's nyere analytiske releases:

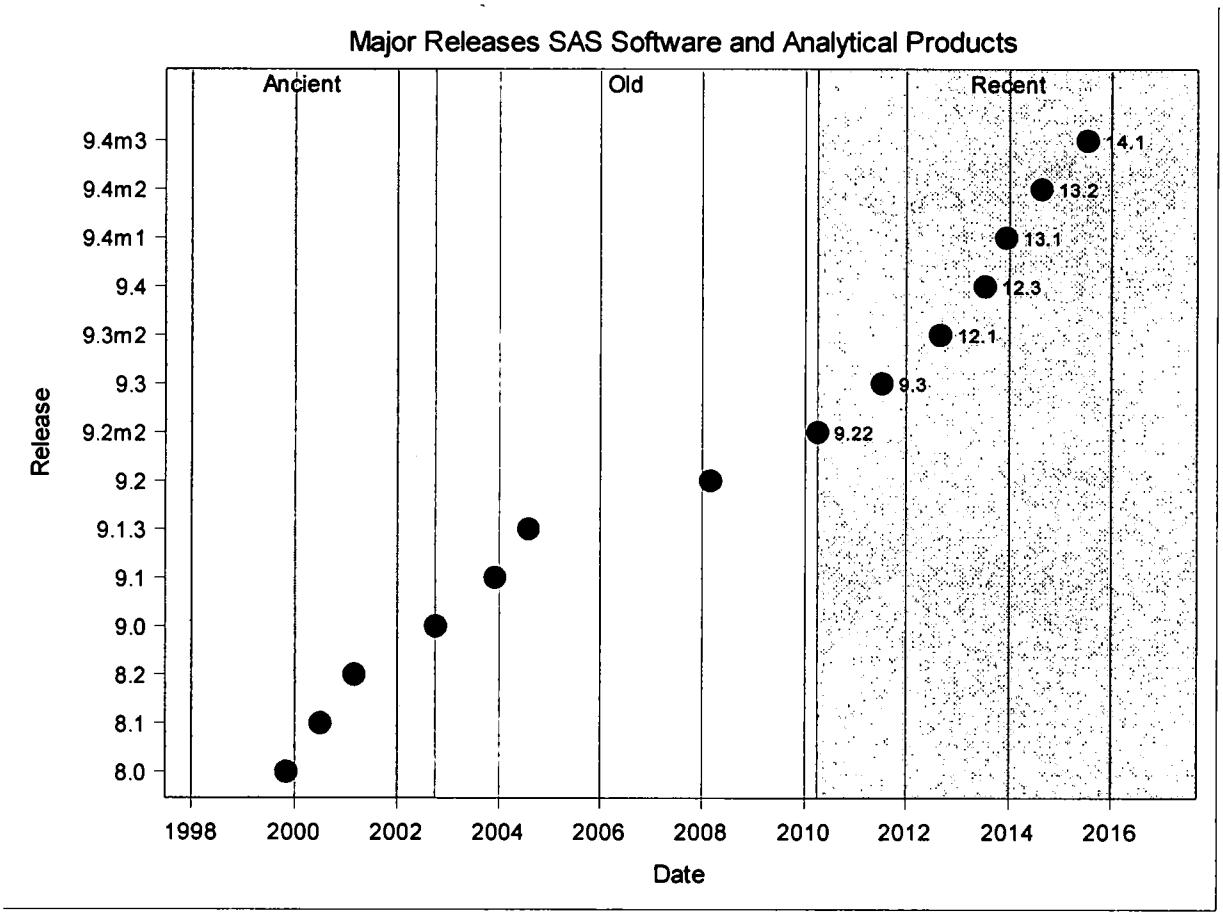
Version 9.4 med Analytical updates 14.1 fra sommeren 2015 er i øjeblikket den aktuelle for universitetsansatte og studerende på fx sasdownload.dk

På en SAS blog,
http://blogs.sas.com/content/iml/2013/08/02/how-old-is-your-version-of-sas-release-dates-for-sas-software/

findes der en oversigt over Analytical updates indenfor de sidste år. Det er selvfølgelig i form af et SAS program, der som output viser en graf, som vist her, idet jeg har tilføjet de nyere opdateringer siden 2013. Den lodrette akse er nok en smule misvisende, da fx springet fra 12.1 til 12.3 udelukkende var vedligeholdelse "maintenance".

I symposieindlægget januar 2015 gennemgik jeg visse af nyhederne i 13.1 og 13.2, som indeholdt en del flere nyheder end en ændring på første decimal ellers skulle antyde. I dette indlæg fokuseres på version 14.1.

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 tilgængelige for alle. Store dele af passagerne på engelsk er direkte citater.



Øget tilgængelighed til SAS

Der blev for et par år siden udviklet et nyt interface til SAS kaldet SAS Studio. Det virker som en kombination af det bedste fra en traditionel kode-SAS tilgang og så den fantastiske editor fra Enterprise Guide.

Desuden er der frigivet en "University Edition", som stilles gratis til rådighed for "alle", der påstår, at de studerer ved et universitet. Den markedsføres som SAS-U. Håbet er, at SAS på den måde kan komme længere "ned" i undervisningssystemet, så allerede high school elever, altså gymnasieelever, kan stifte bekendtskab med de mange muligheder i SAS.

Det bedste ved SAS-U er, at den er udviklet som "virtuel applikation", der kan køres på langt flere platforme end det traditionelle SAS-system. Især kan det uden videre anvendes på en Mac! Det er jo en mulighed, der har været efterspurgt længe i universitetsverdenen, hvor Mac's markedsandel kan være over 50%. Men i det store hele virker SAS-U på samme måde som SAS Studio.

SAS-U kan downloades fra

http://www.sas.com/da_dk/software/university-edition.html

Der kræves en 64 bits maskine og en virtualization software pakke, se

http://www.sas.com/da_dk/software/university-edition.html#m=system-requirements

Som hovedregel går installationen af SAS-U nemt; den meste tid går med at downloade selve SAS-U, som er en fil på knapt 2 Gb. Den skal blot ligge et sted på brugerens harddisk. Når den først er downloaded skal den tilknyttes den virtuelle boks. Her har jeg oplevet, at visse studerende har haft problemer med at få tilordnet deres egne filreferencer korrekt. Moralen er, at man skal gøre hvad der står i vejledningen - bestemt IKKE, hvad man tror, at der står.

En væsentlig ulempe er, at kun SAS-BASE, STAT pakken og IML (matrix-regning) er med i SAS-U. Det skal dog understreges, at selv de underligste esoteriske dele af disse pakker er med i SAS-U. Mit senere eksempel i dette indlæg om PROC SURVEYIMPUTE kan altså udføres med SAS-U, mens eksemplet om PROC RAREEVENTS fra QC-pakken ikke understøttes af SAS-U.

Desuden er dele af ETS (økonometri og tidsrækkeanalyse) pakken inkluderet fra og med sommeren 2014. Det er de dele af ETS, der kan betegnes som "data-scientist" procedurer i modsætning til videnskabelige, professionelle procedures. Det betyder at PROC ESM til forudsigelser og PROC UCM, som anvendes i mit symposieindlæg i år om parkometre, er medtaget i SAS-U. Derimod er fx PROC VARMAX og PROC X12 ikke medtaget.

SAS-U afvikles som en applikation i en internet browser. Det betyder, at filer på harddisken på brugerens egen PC skal tilgås på en anden måde en normalt via "delte filer" i den virtuelle boks; men svært er det ikke.

En yderligere ulempe er, at SAS-U i praksis ikke kan håndtere store datamængder med titusindvis af observationer med hundredevis af variable. Det skyldes selvfølgelig, at tilgangen via en internetbrowser ikke kan optimere adgangen til CPU og RAM på samme effektive måde som en traditionel SAS-installation. Et eksempel med den danske del af PISA undersøgelsen med knapt 8000 observationer af 738 variable kører dog uden problemer, men en analyse af det samlede PISA materiale er i praksis umuligt.

SAS og Big Data

Listen over HP (High Performance) procedurer, som kopierer de gængse SAS procedurer, udvides stadigt. De indeholder ikke de mange grafiske faciliteter, for man kan jo ikke tegne millionvis af punkter i et diagram. Men de regner hurtigt, og de udnytter maskinfigurationen fuldt ud. Fx kan de regne multi-threadet, hvis maskinen indeholder flere processorer. Det nye i version 9.4 er, at de også stilles til rådighed for almindelige SAS brugere i simple PC installationer. Derved kan visse ekstra raffinementer udnyttes, og programmer kan afprøves før produktionskørsler på fjerne servere med stor

maskinkraft. Det kræver specielle licenser at udnytte faciliteterne til distribueret kørsel af SAS-programmer.

Disse muligheder for analyser af Big data virker noget fjerne for an almindelig universitetsansat eller student. Men nogen skulle arbejde for at forskere fik mulighed for at afprøve mulighederne i praksis, fx på forskermaskinerne i Danmarks Statistik, hvor datamængderne kan være betydelige.

Ud fra navnene på procedurerne kan man ved a fjerne præfikset HP se hvilke procedurer, der indtil videre findes i en High Performance. Der er kommer nye high performance procedurer i alle versioner - også til 14.1, så den samlede liste er lige pt:

STAT	ETS
The HPCANDISC Procedure	The HPCDM Procedure
The HPFMM Procedure	The HPCOPULA Procedure
The GAMPL Procedure	The HPCOUNTREG Procedure
The HPGENSELECT Procedure	The HPPANEL Procedure
The HPLMIXED Procedure	The HPQLIM Procedure
The HPLOGISTIC Procedure	The HPSEVERITY Procedure
The HPNLMOD Procedure	
The HPPLS Procedure	
The HPPRINCOMP Procedure	
The HPQUANTSELECT Procedure	
The HPREG Procedure	
The HPSPLIT Procedure	

Whats New in 14.1 sammenholdt med 13.2

I dette afsnit citeres nyhedsoplistningerne for de to pakker STAT og ETS. I mit symposieindlæg for 2015 var der en tilsvarende liste for 13.2. De i praksis halvårlige opdateringer as SAS Analytical Updates kan installeres ved at følge med de løbende TS#M# serviceopdateringer.

Nyheder i SAS/STAT 14.1 sammenholdt med 13.2

To nye procedurer:

> **SURVEYIMPUTE Procedure (se eksempel senere)**
> The SURVEYIMPUTE procedure imputes missing values of an item in a sample survey by replacing them with observed values from the same item.

GAMPL Procedure

The GAMPL procedure is a high-performance procedure that fits generalized additive models by penalized likelihood estimation.

Following are some highlights of the enhancements in SAS/STAT 14.1:

- The BCHOICE procedure allows varying numbers of alternatives in choice sets for logit models.
- Exact mid-p, likelihood ratio, and Wald modified confidence limits are available for the odds ratio produced by the FREQ procedure.
- The GLIMMIX procedure provides the multilevel adaptive Gaussian quadrature algorithm of Pinheiro and Chao (2006) for multilevel models, which can greatly reduce the computational and memory requirements for these models with many random effects.
- The GLMSELECT procedure supports the group LASSO method.
- The IRT procedure fits generalized partial credit models.
- The LIFETEST procedure performs nonparametric analysis of competing-risks data.
- The LOGISTIC procedure fits an adjacent-category logit model to ordinal response data.
- The MCMC procedure adds an ordinary differential equation (ODE) solver and a general integration function, enabling the procedure to fit models that contain differential equations (for example, PK models) or models that require integration (for example, marginal likelihood models).
- The NPAR1WAY procedure performs stratified rank-based analysis for two-sample data.
- The POWER procedure supports Cox proportional hazards regression models.

Nyheder i SAS/ETS 14.1 sammenholdt med 13.2

New procedure **PROC X13**, incorporating the SEATS method into the X-12-ARIMA seasonal adjustment program

New features have been added to the following SAS/ETS components:

- COUNTREG procedure
- HPCOUNTREG procedure
- HPPANEL procedure
- MODEL procedure

- PANEL procedure
- QLIM procedure
- SSM procedure
- VARMAX procedure

fx: for VARMAX procedure: Kraftige forbedringer af ARCH og Cointegration

What's New in SAS/QC 14.1: New RAREEVENTS Procedure (Experimental)

The RAREEVENTS procedure produces control charts for rare events. A rare event is one that occurs infrequently, with a low probability. A rare events chart is better suited than traditional control charts to detecting changes in the frequency of low-probability events.

Jordskælv i Oklahoma

I dette afsnit anvendes den nye procedure PROC RAREEVENTS på et datasæt, der bla indeholder eksakt tidspunkt, på sekundniveau, samt styrke for alle jordskælv i Oklahoma. Datasættet er downlaoded fra

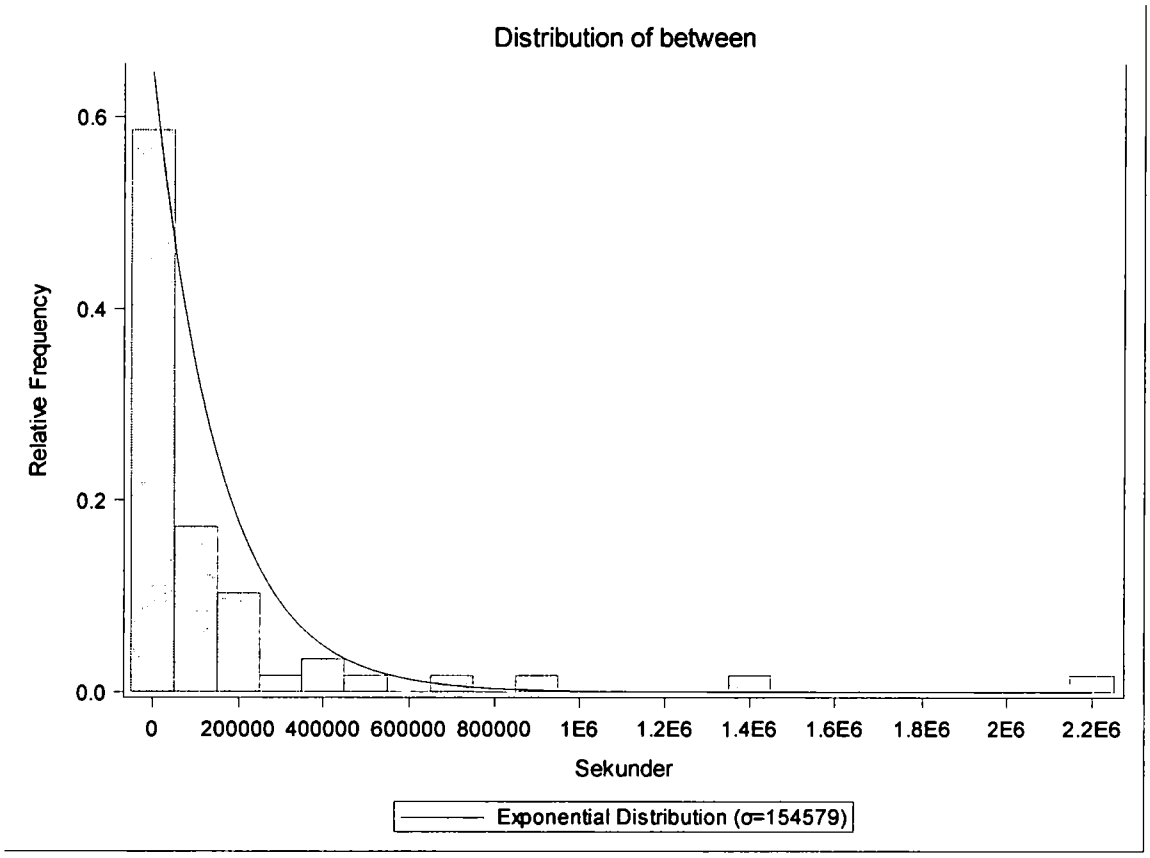
htttp://earthquake.usgs.gov

Nogle påstår at den udbredte udvinding af olie med den nye "fracking" metode leder til flere jordskælv. Dette kan let illustreres med PROC RAREEVENTS.

I det første program analyseres data fra perioden 1. jannuar 2010 til 1. juli 2014. I den periode var der 58 jordskælv af en styrke over 3.5.

```
PROC RAREEVENTS DATA=wrk8.Oklahoma;
   COMPARE BETWEEN/NBINS=45;
   CHART BETWEEN * time/ALPHALPL=0.05 ALPHAUPL=0.05
       OUTLIMITS=L;
   LABEL BETWEEN = 'Sekunder';
   ID time;
   where year(datepart(time))>2010 and
       time<'01jul2013:00:00:00'dt and mag>3.5;
RUN;
```

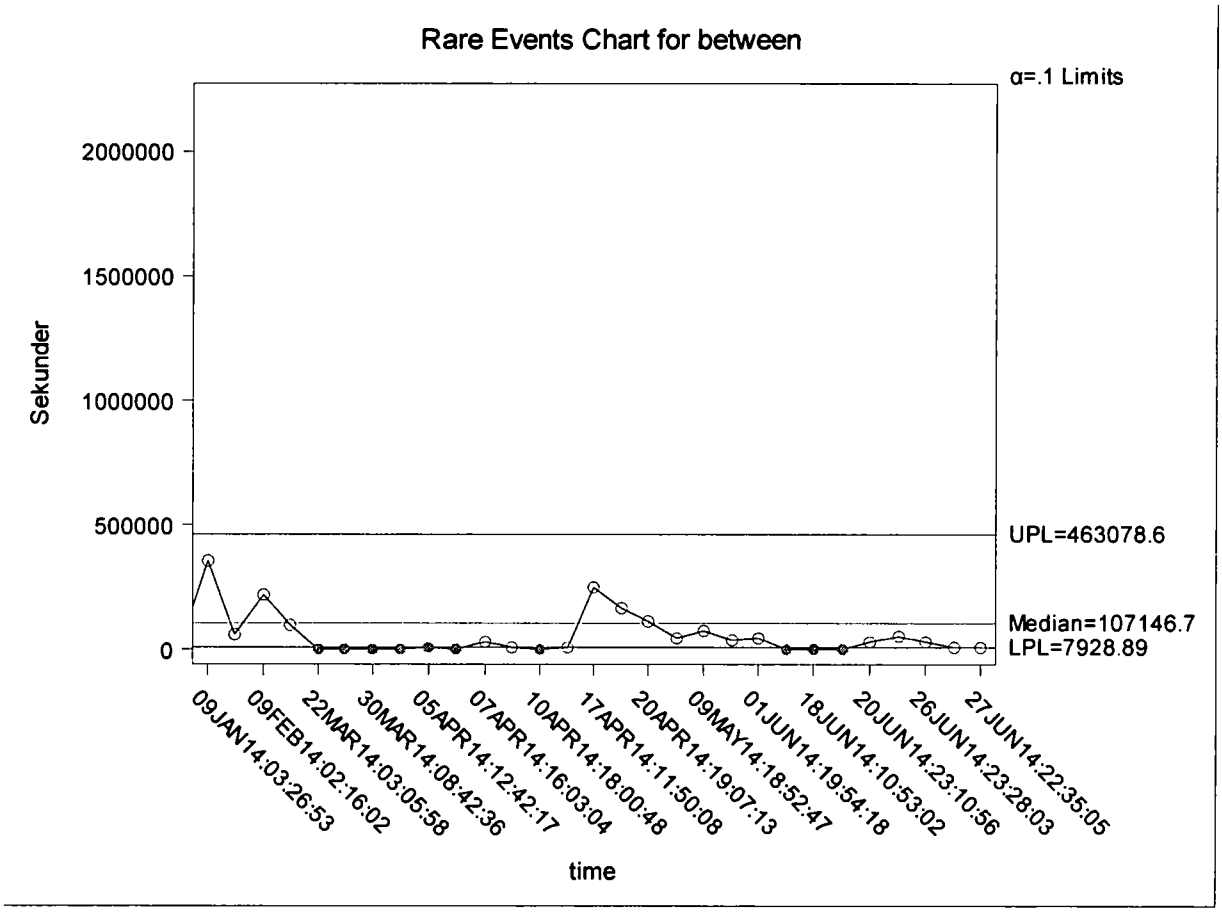
Blandt outputtet er der et histogram over ventetiderne mellem to jordskælv af styrke over 3.5:



Desuden er der et plot over ventetiderne mellem to jordskælv afsat mod tidspunktet for jordskælvet. På dette plot er der indtegnet sikkerhedsgrænser omkring den forventede ventetid, så man kan se om ventetiden er usædvanlig kort eller lang.

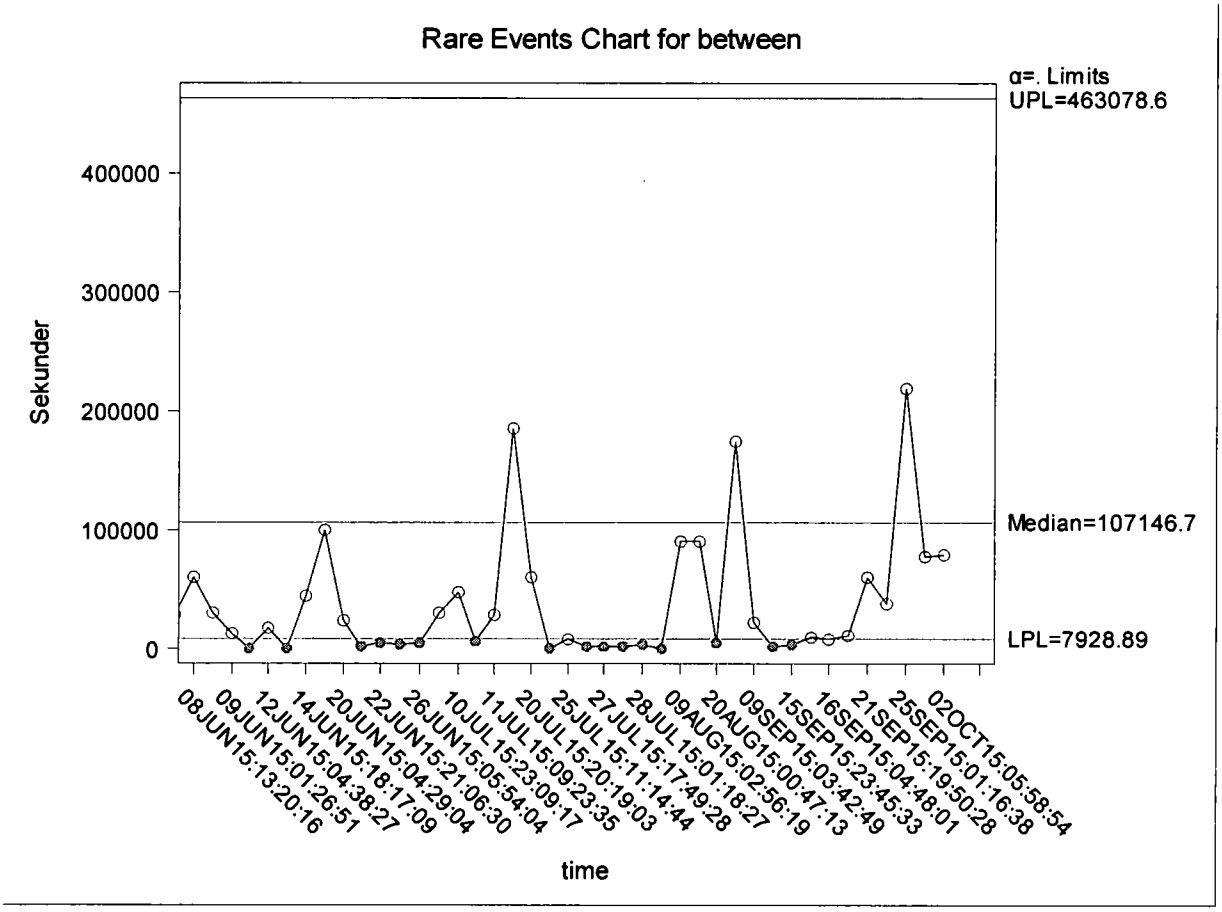
Tidsenheden er sekunder, så tallene er store. Grænserne for konfidensgrænserne er bestemt ud fra en konfidensgrad på 90%, dvs at det må forventes at 5% ligger under og 5% over konfidensbåndet. På plottet er angivet de 29 seneste ud af de i alt 58 jordskælvskælv, hvoraf 10 ligger under konfidensbåndet.

Disse grænser gemmes i datasættet kaldet L ved optionen OUTLIMITS=L. I det næste program anvendes i det næste program som grænser for de tilsvarende data for perioden 1/7 - 2014 til ca 1/10 2015.




```
PROC RAREEVENTS DATA=wrk8.Oklahoma LIMITS=L;
   COMPARE BETWEEN/NBINS=45;
   CHART BETWEEN * time/ALPHALPL=0.05 ALPHAUPL=0.05
        OUTTABLE=ASDFG;
   LABEL BETWEEN = 'Sekunder';
   ID time;
   where time>'01jul2014:00:00:00'dt and mag>3.5;
RUN;
```

I denne periode er der 127 jordskælv, så plottet over ventetiderne er delt over 3 plots på hver ca 43 observationer. De seneste af de tre plots vises her:



Ud af de 43 punkter ses de 15 at ligge under konfidensbåndet, hvor der egentlig kun burde ligge 5% dvs kun 2- 3 af punkterne.

PROC SURVEYIMPUTE

PROC SURVEYIMPUTE implements imputation techniques that do not use explicit models: hot-deck imputation, cold-deck imputation, and fractional imputation.

Hot-deck imputation is the most commonly used imputation technique for survey data. A donor is selected for a recipient unit, and the observed values of the donor are imputed for the missing items of the recipient.

Fractional hot-deck, also known as fractional imputation (FI), is a variation of hot-deck imputation in which one missing item for a recipient is imputed from multiple donors. Each donor donates a fraction of the original weight of the recipient such that the sum of the fractional weights from all the donors is equal to the original weight of the recipient. For fully efficient fractional imputation (FEFI), all observed values in an imputation cell are used as donors for a recipient unit in that cell.

PROC SURVEYIMPUTE also produces replicate weights that can be used with any survey analysis procedure in SAS/STAT to estimate both the sampling variability and the imputation variability.

Et eksempel på brug af PROC SURVEYIMPUE

For at demonstrere de mange muligheder med PROC SURVEYIMPUTE skal der anvendes data fra en stikprøve, der indsamlet med stratifikation, med clustre og hvor replikationsvægte er stillet til rådighed. Sådanne datasæt findes der kke mange af Danmark; jeg kender kun PISA, der er indsamlet efter internationale forskerstandarder.

Der anvendes den nyeste danske runde af PISA. blandt de mage variable i datasættet anvendes visse spørgsmål om standarden i barnets hjem. Disse variable er indsamlet ved et spørgeskema, hvor ikke alle har besvaret alle spørgsmålene.

For at forklare hvad analysen går ud, præsenteres slutresultatet først. Den følgende tabel viser resultatet af en logistisk regression hvor svaret "Yes" til spørgsmålet "Er der poesi i hjemmet?" Den viste tabel inkluderer imputation med PROC SURVEYIMPUTE og variansestimation baseret på replikationsvægte.

Analysis of Maximum Likelihood Estimates

Parameter		*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Label*
Intercept		-1.4805	0.3565	-4.15	<.0001	Intercept: ST26Q08=No
ST26Q02	*Yes*	0.0200	0.0520	0.39	0.7010	Possessions - own room Yes
ST26Q06	*Yes*	0.1436	0.3960	0.36	0.7178	Possessions - Internet Yes
ST26Q13	*Yes*	0.1188	0.1022	1.16	0.2484	Possessions - dishwasher Yes
ST26Q14	*Yes*	0.2563	0.0988	2.60	0.0112	Possessions - <DVD> Yes
ST04Q01	*Male*	-0.1450	0.0531	-2.73	0.0077	Gender Male

NOTE: The degrees of freedom for the t tests is 80.

Resultatet er baseret på to procedurekald. Først kaldet af PROC SURVEYIMPUTE, hvor metoden angives til FEFI.

```
proc surveyimpute data=wrk8.dk_reduce
        method=FEFI varmethod=jk;
   weight W_FSTUWT;
   repweights W_FSTR: ;
   class ST26Q08 ST26Q02 ST26Q06 ST26Q13 ST26Q14 ST04Q01;
   var ST26Q08 ST26Q02 ST26Q06 ST26Q13 ST26Q14 ST04Q01;
   output out=Imputed;
run;
```

Variablen W_FSTUWT i WEIGHT statementet er enkelte respondenters vægt i analysen; dvs i praksis er det en vægt variabel for indvandrerstatus, da der oversamplet blandt skoler med en høj indvandrerandel.

Replikationsvægtene angives i en REPWEIGHT statement. Kolonet i W_FSTR: dækker over, at der anvendes 80 variable W_FSTR1-W_FSTR80.

I outputtet fra PROC SURVEIMPUTE findes en oversigt over manglende værdier.

Group	*ST26Q08*	*ST26Q02*	*ST26Q06*	*ST26Q13*	*ST26Q14*	*ST04Q01*	*Freq*
1	X	X	X	X	X	X	7153
2	X	X	X	X	.	X	10
3	X	X	X	.	X	X	11
4	X	X	X	.	.	X	1
5	X	X	.	X	X	X	7
6	X	.	X	X	X	X	9
7	.	X	X	X	X	X	151
8	.	X	X	X	.	X	5
9	.	X	X	.	X	X	14
10	.	X	X	.	.	X	3
11	.	X	.	X	X	X	1
12	.	X	.	.	X	X	1
13	.	X	.	.	.	X	8
14	.	.	X	X	X	X	2
15	.	.	X	X	.	X	1
16	.	.	X	.	X	X	1
17	.	.	X	.	.	X	2
18	.	.	.	.	.	X	101

Det er oplagt, at det ikke giver mening at imputere de 5 øvrige variable om hjemmets stand udelukkende ud fra barnets køn. Det er også tvivlsomt om det kan siges noget om forekomsten af poesi i hjemmet ud fra en oplysning om hjemmet har en opvaskemaskine.

Datasættet indeholder heldigvis mange flere spørgsmål om hjemmets stand end de få, der er vist her. Imputationerne, der ligger bag ved det viste slutresultat, er selvfølgelig baseret på en langt større informationsmængde ud fra alle tilgængelige variable.

Ud fra de knapr8000 observationer i det oprindelige datasæt dannes datasættet IMPUTE, der indeholder knapt 56.000 observationer. Datasættet IMPUTE er dannet ud fra hver a 80 replikationsvægte, der anvendt som vægte i 80 forskellige Fully Effficient Fractional Imputations. I det nye datasæt er vægtene og replikationsvægtene justeret svarende til det højere antal "observationer" i et nye datasæt.

Den afsluttende logistiske regression udføres med PROC SURVEYLOGISTIC:

```
proc surveylogistic data=Imputed varmethod=jk;
    class ST26Q02 ST26Q06 ST26Q13 ST26Q14
        ST04Q01/descending;
    model ST26Q08(event='Yes')=ST26Q08 ST26Q02 ST26Q06
        ST26Q13 ST26Q14 ST04Q01;
    weight W_FSTUWT;
    repweights W_FSTR: ;
run;
```

Ligestilling, religion og nationalitet

af

Niels Kærgård, Institut for Fødevare- og Ressourceøkonomi, KU,

Peter Lüchau, Institut for Kultur og samfund, AU,

Anders Milhøj, Økonomisk Institut, KU

Indledning

Danmark har i de sidste årtier fået en række indvandrergrupper, for hvem ligestilling mellem kønnene ikke er så naturlig, som for de fleste moderne danskere. Er det naturligt, at kvinderne er hjemmegående og passer børnene? Har mændene mere ret til jobbene end kvinderne? En del debattører har henført disse holdninger til indvandrernes religion, specielt islam. Andre tillægger traditioner i hjemlandet større betydning. Det er det, der skal ses på i det følgende.

Grundmaterialet er en større undersøgelse af værdier og holdninger blandt udlændinge og danskere lavet af Integrationsministeriets Tænketank sammen med sociologerne Peter Gundelach og Esther Nørregård-Nielsen i 2007 (Tænketank, 2007 og 2007a). Her blev 6 indvandrergrupper (indvandrere fra Tyrkiet, Vestbalkan, Irak, Iran, Pakistan og Vietnam), 2 grupper af efterkommere (fra Tyrkiet og Pakistan) og en gruppe danskere spurgt om tro, holdninger og værdier. I alle grupper er der fra starten udvalgt ca. 1000 tilfældige personer, og det giver så med en svarprocent på omkring 50 følgende antal interview, se tabel 1.

Tabel 1: Antal interview og svarprocent

	Antal interview	Svarprocent
Indvandrere		
Tyrkiet	503	48 %
Vestbalkan	505	54 %
Irak	513	60 %
Iran	505	61 %
Vietnam	506	51 %
Pakistan	434	41 %
Efterkommere		

Tyrkiet	504	52 %
Pakistan	488	52 %
Danskere	520	67 %

Kilde: Tænketanken, 2007, side 33.

Der blev bl.a. spurgt om hvilen religion, de tilhørte, og om de var religiøse med fire svarmuligheder. Vedrørende ligestilling blev grupperne bl.a. spurgt om mænd har mere ret til arbejde end kvinder, hvis der ikke er job til alle. Den dominerende religion i alle andre grupper end vietnamesere og danskere er islam. Så der blev i rapporten specielt set på om mere religiøse muslimer havde en anden holdning end mindre religiøse muslimer. I den forbindelse ses i rapporten på følgende tabel, her tabel 2, og det konkluderes umiddelbart, at oprindelsesland synes at have stor betydning, mens religiøsitet har mindre betydning. En lignende konklusion drages i Kærgård (2015). Tallene er imidlertid ikke her analyseret ved hjælp af matematisk-statistiske metoder.

Tabel 2: Andel af muslimer, der er uenig i, at mænd har mere ret til arbejde end kvinder, hvis der ikke er job til alle, fordelt efter graden af religiøsitet

	Religiøs/meget religiøs	Ikke særlig religiøs/slet ikke religiøs
Indvandrere		
Tyrkiet	40 %	46 %
Vestbalkan	73 %	75 %
Irak	53 %	69 %
Iran	70 %	76 %
Pakistan	75 %	69 %
Efterkommere		
Tyrkiet	71 %	74 %
Pakistan	80 %	79 %

Kilde: Tænketanken, 2007, side 117.

Sådanne metoder blev brugt i Kærgård & Milhøj (2015), men her alene baseret på antal i de enkelte celler beregnet ud fra de marginale hyppigheder. Siden har vi fået adgang til grunddata og beregningerne skal i dette papir gentages på grundlag af de direkte observerede svar i de enkelte celler.

Analysen vil - helt som i Kærgård & Milhøj (2015) - blive gennemført ved hjælp af loglineære modeller for kategoriserede data, se fx Andersen(1997). I data er der tre inddelingskriterier:

Oprindelsesland med 7 kategorier, idet efterkommere er en "nation" for sig. Variablen **"Land"**.

Religiøsitet med 2 kategorier der afspejler høj/lav grad af religiøsitet. Variablen **"R"**.

Mænd forret til arbejdet med 2 kategorier Ja/Nej. Variablen **"M"**.

En log-lineær model for disse tre variable parameteriseres ved cellesandsynlighederne:

$$P(Land = l, R = r, M = m)$$

$$= exp(\tau_{l,r,m}+\tau_{l,r} + \tau_{l,m} + \tau_{r,m}+\tau_l + \tau_r + \tau_m + \tau_0)$$

for en værdi, l, af variablen land, en værdi, r, af variablen for religiøsitet og en værdi, m, for variablen, der angiver svaret på spørgsmålet om mænds forret til arbejde. Alle τ-er i dette udtryk summerer til nul, når der summeres over et af indeksene i fodtegnene. Formlen kan genskabe enhver empirisk hyppighed, da der er lige mange frie parametre, τ-er , som frit varierende celler i datamaterialet. Det er den mættede model. I dette datamateriale er der i alt svar fra 3299 personer, og den svagest repræsenterede celle, dvs kombination af land, religiøsitet og synspunkt, har 13 observationer. Det betyder, at de asymptotiske teknikker, der anvendes ved estimation og tests, må antages at være fuldt tilfredsstillende.

For de fleste oprindelseslande er der uafhængighed mellem graden af religiøsitet og holdning til mænds forret til arbejdet. For Tyrkiet ses en svag sammenhæng, men med en p-værdi på 3.4% er afhængigheden signifikant, men borderline. For Irak er det meget signifikant, at meget religiøse i højere grad end mindre religiøse mener, at mænd har forret til arbejde.

Tabel 3: Andelen med holdnignen "Mænd har forret til arbejde opdelt efter graden af religiøsitet. P-værdien er test for om disse andele land for land er ens

	Religiøs/meget religiøs	Ikke særlig religiøs/slet ikke	

		religiøs	
Indvandrere			
Tyrkiet	54%	44 %	p = 0.034
Vestbalkan	22 %	17 %	p = 0.251
Irak	42 %	25 %	p = 0.0002
Iran	23 %	17 %	p = 0.187
Pakistan	18 %	22 %	p = 0.513
Efterkommere			
Tyrkiet	20 %	18 %	p = 0.668
Pakistan	11 %	11 %	p = 0.610

I denne analyse er det naturligt at inddrage respondentens køn. I en model med i alt fire variable bliver fjerde ordens vekselvirkningen insignifikant, mens to af tredje ordens vekselvirkningerne, hvori variablen land indgår, er signifikante. Det er derfor naturligt igen at opdele analysen i oprindelseslandene hver for sig.

For alle syv lande gælder, at tredjeordens vekselvirkningen er insignifikant. Desuden er sammenhængen mellem graden af religiøsitet og køn signifikant for alle lande undtagen Vestbalkan (dog borderline, p = 6%) og mere bemærkelsesværdigt også efterkommere fra Tyrkiet, p = 76%.

Kun for efterkommere fra Tyrkiet og indvandrere fra Iran er der en sammenhæng mellem køn og holdningen til mænds forret til arbejde. For efterkommere fra Tyrkiet er sammenhængen dog svag, p = 3.8%, i retning af, at mænd ikke har forret. for Iran går sammenhængen stærkt den modsatte vej, idet kvinder i højere grad mener, at mænd har forret til arbejdet, p = 0.6%.

Sammenligning af generationer

Kærgård og Milhøj (2015) betragtede også forskelle mellem generationer for indvandrere fra Tyrkiet og Pakistan, hvor både første generationsindvandrere og efterkommere er repræsenteret. Med de faktiske tal er bliver resultaterne stort set de samme som konklusionen ud fra de skønnede tal, der blev anvendt i Kærgård og Milhøj (2015).

For tyrkere er det meget signifikant, p = 0.01%, at efterkommere er mere religiøse, 71%, end første generations indvandrere, 60%. For pakistanere findes den omvendte effekt, med 76% religiøse blandt indvandrere mod 84%+ blandt første generations indvandrere; se tabel 4. Effekten bland pakistanere er ligeledes signifikant, p = 0.8%.

Spørgeskemaundersøgelsen indeholdt også en variabel om hvorvidt respondenten følte sig mere, mindre eller uændret religiøs i forhold til for 3 år siden. Denne variabel er afhængig af graden af religiøsitet, hvilket vel nærmest kan opfattes som en

definitionssag. Men den er også afhængig af genrationen, som vist i følgende tabeller, med p-værdier under 0.1%.

Tabel 4: Sammenhæng mellem generation og udvikling i religiøsitet for Tyrkiet

	Mere religiøs	Mindre religiøs	Uændret grad af religiøsitet	Total
Efterkommer	163 33.13	30 6.10	299 60.77	492
Indvandrer	95 19.08	33 6.63	370 74.30	498
Total	258	63	669	990

Tabel 5: Sammenhæng mellem generation og udvikling i religiøsitet for Pakistan

	Mere religiøs	Mindre religiøs	Uændret grad af religiøsitet	Total
Efterkommer	204 42.06	51 10.52	230 47.42	485
Indvandrer	146 33.72	28 6.47	259 59.82	433
Total	350	79	489	918

For holdningsspørgsmålet om mænds forret til arbejde er fordelingen mellem tyrkiske og pakistanske indvandrere ligeledes forskellig mellem første generations og efterkommere. Blandt tyrkere er der en klar overvægt af holdningen om, at mænd har forret til arbejde blandt første generations indvandrere, hele 50%, men kun 19% af efterkommerne deler dette synspunkt. Bandt pakistanere er det kun 19% blandt første genrations indvandrere og 13% blandt efterkommere. Disse forskelle er stærkt signifikante med p-værdier under 1% for begge lande. Efterkommerne nærmer sig altså danskerne med hensyn til holdningen til kønslig ligestilling. Dette helt uanset at efterkommerne er mere religiøse end førstegenerations indvandrerne og stadig forøger deres religiøsitet.

Sammenligning af religioner

Disse analyser vedrører alene muslimer, der er den helt dominerede religion i alle de interviewede grupper bortset fra danskerne og indvandrerne fra Vietnam, og de var ikke med i analyserne ovenfor. Analysen ovenfor vedrører altså alene religiøse muslimer overfor ikke-religiøse muslimer. Det ville imidlertid også være interessant at se på de forskellige religioners indflydelse.

Kun for Vestbalkan og Vietnam er der mere end én religiøs gruppe af væsentlig størrelse. For Vestbalkan er 69 % muslimer, 14 % ortodoks kristne og 6 % romersk katolske. For Vietnam er 43 % buddhister og 37 % romersk-katolske.

For disse lande hver for sig gennemføres en tilsvarende analyse med tre inddelingskriterier:

Tro med 4 kategorier, muslimer, ortodoks kristne, ikke-religiøse og romersk katolske. Variablen **"Tro".**

Religiøsitet med 2 kategorier der afspejler høj/lav grad af religiøsitet. Variablen **"R".**

Mænd forret til arbejdet med 2 kategorier Ja/Nej. Variablen **"M".**

En log-lineær model for disse tre variable parameteriseres ved cellesandsynlighederne:

$$P(Tro = t, R = r, M = m)$$

$$= exp(\tau_{t,r,m}+\tau_{t,r} + \tau_{t,m} + \tau_{r,m}+\tau_t + \tau_r + \tau_m + \tau_0)$$

for en værdi, t, af variablen tro, en værdi, r, af variablen for religiøsitet og en værdi, m, for variablen, der angiver svaret på spørgsmålet om mænds forret til arbejde. Alle τ-er i dette udtryk summerer til nul, når der summeres over et af indeksene i fodtegnene. Formlen kan genskabe enhver empirisk hyppighed, da der er lige mange frie parametre, τ-er , som frit varierende celler i datamaterialet. Det er den mættede model.

For Vestbalkan er tredjeordens vekselvirkningen, $\tau_{t,r,m}$, ikke signifikant, og desuden er de to anden ordnes vekselvirkninger mellem hhv tro og holdningen til mænds forret

til arbejdet, og mellem graden af religiøsitet og holdningen til mænds forret til arbejde, dvs parametrene $\tau_{t,m}$ og $\tau_{r,m}$, ikke signifikante.

Modellen uden disse parametre accepteres ved et test mod den mættede model med p = 77%. Den eneste sammenhæng mellem de tre variable er altså en sammenhæng mellem tro og graden af religiøsitet. Et test for uafhængighed mellem disse to variable forkastes for Vestbalkan med p = 0.029. Sammenhængen beskrives i den følgende tabel, der viser, at muslimer i højere grad, 46%, betragter sig selv som troende end de to kristne grupper, hhv 34% og 30%.

Tabel 6: Sammenhængen mellem tro og religiøsitet for indvandrere fra Vestbalkan

Tro	Religiøsitet		
	Ikke religiøs	religiøs	Total
Romersk katolsk	21	11	32
	65.63	34.38	
Ortodoks kristen	50	21	71
	70.42	29.58	
Muslim	187	157	344
	54.36	45.64	
Total	258	189	447

Også for Vietnam viser en tilsvarende analyse, at der er uafhængighed mellem troen og holdningen til mænds forret til arbejdet. Derimod er der også her afhængighed mellem troen og så graden af religiøsitet, idet de romersk katolske er mere religiøse end buddhisterne. Der er også en sammenhæng mellem graden af religiøsitet og så holdningen til mænds forret til arbejdet med en p-værdi under 1%.

Sammenfatning

Hvor det i den offentlige debat i Danmark og mange andre lande betragtes som givet, at ligestillingsproblemer i en række indvandrergrupper kan henføres til stærkt

muslimske miljøer, så peger denne undersøgelse på, at en forbindelse mellem religion, religiøsitet og holdninger til ligestilling mellem kønnene er usikker og insignifikant.

Der er derimod klare forskelle i holdningen til ligestilling mellem de forskellige indvandrergrupper efter, hvilket land de kommer fra. Oprindelsesland er en langt mere betydningsfuld forklarende variabel end religion og graden af religiøsitet.

Traditioner i hjemlandet synes således en central variabel, og man kan ikke udelukke, at religion blandt tidligere generationer kan have spillet en rolle i skabelsen af sådanne traditioner. Der kan være meget lange forsinkelser fra religion til holdninger og adfærd. North, Orman & Gwin (2013) finder f.eks. vedrørende korruption, at religionen for hundrede år siden har mere forklaringskraft end aktuel religion.

Men hvis der er en forbindelse fra religion til holdninger vedrørende f.eks. ligestilling, så er forbindelsen i hvert fald langt fra simpel og direkte.

Litteratur

Andersen, Erling B. (1997), *Introduction to the Statistical analysis of Categorical Data*, Springer

Kærgård, Niels (2015), *From Religion to Economic Acting*, Paper for annual meeting in The Association of the Study of Religion, Economics and Culture, Boston, March 20-22, 2015.

Kærgård, Niels & Anders Milhøj (2015), Religion og oprindelseslands betydning for indvandrers holdninger, *Symposium i anvendt statistik 2015*, Danmarks Statistik, København.

North, Charles M., Wafa Hakim Orman & Carl R. Gwin (2013), Religion, Corruption and the Rule of Law, *Journal of Money, Credit and Banking*, Vol. 45.5, side 757-779.

Tænketanken (2007), *Værdier og normer – blandt udlændinge og danskere*, Ministeriet for flygtninge, indvandrere og integration, København.

Tænketanken, Peter Gundelach & Ester Nørregård-Nielsen (2007a), *Etniske gruppers værdier - Baggrundsrapport*, Ministeriet for flygtninge, indvandrere og integration, København.

Difficulty and Grade Distribution Analysis of Compulsory Courses at the Economy Program at the University of Copenhagen[1]

Sara Armandi

1 Introduction

Universities have always had great impact on a nation's innovation and development. In a desire to increase the contribution from universities to the economy, the importance of universities in the political arena has grown. This has enhanced the attention on the spheres of the universities.

At present time, the university system in Denmark is initiating a new reform. The *Study Progress Reform*, as it is called, was passed by the Danish Parliament in 2013. It has the agenda to get students through the educational system as fast as possible and into the job market in order to increase the number of tax payers and decrease the expenses on education. The universities are facing a dilemma; on one hand they are granted for each student completing a degree, while on the other hand, if the universities do not succeed in decreasing the total study duration of its students, they will lose large amounts of money as fines will be distributed. Even though it may be appealing for the universities to expel low ability and lazy students who might increase the total study duration, it could turn out to be unprofitable as some students might eventually graduate even though they do not have the best prerequisites.

Large amounts of data are collected for purely administrative purposes in the field of education. The potential insights stored in the data could help the universities in understanding the educational process of the students.

By realizing which courses are the most difficult ones, and by examining whether the ability of students affects the study progress, it is possible to provide explicit information

[1] This paper is a summary of the author's master's thesis from January 2016

to the universities on how they can improve the performances and the study progress of their students and, eventually, increase earnings.

Classical test theory (CTT) is one possible way of measuring the difficulty of a course and the abilities of students. However, one of the main drawbacks of CTT is that student characteristics and characteristics of courses cannot be separated. This is not very convenient, as it is troublesome to compare students who partake in different courses and difficult to compare courses who are obtained using different groups of students. In order to overcome these weaknesses *Rasch measurement theory* is used to quantify the properties.

1.1 Rasch and Item Response Theory

Rasch measurement theory is one example of a test theory that overcomes some of the shortcomings of CTT. The first steps toward the Rasch measurement theory was made by the Danish mathematician, Georg Rasch, in the early 1950s. In 1952 Rasch transferred the method to a project for the Danish military where he was to analyze intelligence test data. During his work he developed a model which had the property of separability between item and subject parameters [Rasch, 1960]. This model is now known as the Rasch Model [Andersen and Olesen, 2001].

A separate but highly related line of development in measurement theory is traced to the American statistician Allan Birnbaum. Independently of the work by Rasch he contributed with some chapters about *item response theory* (IRT) to Lord and Novick's "*Statistical Theories of Mental Test Scores*". The models within IRT specify a relationship between an observable subject performance and an unobservable latent ability, which is assumed to underlie the test results. This is identical to what the Rasch model does.

Especially in the 1960s and 1970s Rasch analysis and IRT have been developed, improved, and have rapidly become mainstream as the theoretical basis' for measurement.

The two separate lines of measurement have merged and today they are collectively referred to as IRT [Embretson and Reise, 2000]. These models are capable of evaluating both student ability and item properties, such as item difficulty and discrimination, they are highly useful when trying to answer the research questions stated above.

This paper is the first to implement IRT on an educational program instead of merely using a single test. The usual practice is to consider a number of questions or items which together constitutes a test. An example of such a test could be a mathematics test with different questions within subtraction and multiplication. Hence, in this paper the test is more general and not as specific as usual. The items considered in this paper are 17 end-of-course exams which together form the test, at the bachelor's program in economics at the University of Copenhagen.

The focus of this paper is to investigate and identify the most difficult courses for the students. It will be examined whether different courses are equally easy to pass and whether the grading scale is used consistently across courses. The study progress of the students will be examined by looking at the dropout rates as well as the completion time.

2 Theory

In the following sections the fundamental features of the Rasch model and some of its extensions are presented and the assumptions of the models are discussed.

2.1 Ability

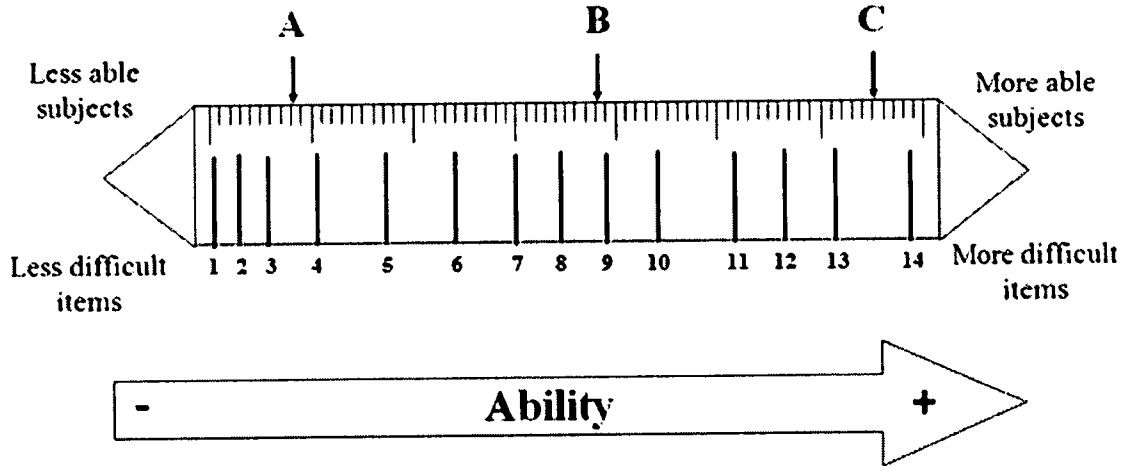
A primary goal of educational measurement is the determination of how much of a latent trait a student possesses. A latent trait, or a latent variable, is an unobservable entity that influences observable variables such as test scores or item responses [Embretson and Reise, 2000]. In IRT the term "ability" is used to refer to the latent variable [Baker, 2001].

To measure the level of ability that a student possesses, an arbitrary underlying ability scale is defined. This ability scale has a midpoint of zero and a range from negative

infinity to positive infinity. Using this scale the amount of ability a student possesses can be determined, and abilities of different students can be compared. The main purpose of IRT is to place items and students on the same measurement scale.

Figure 1 shows three students (*A*, *B*, and *C*) who are all located along the same measurement scale according to their ability. Students with a low ability are located on the left of the scale while students with a higher ability are located on the right. Items can be located on the same scale according to their difficulty. These are indicated by the thick lines numbered from 1 to 14 in Figure 1. Less difficult items are located on the left of the scale while the most difficult items are located on the right. In general, items can be solved by students who have an ability which is higher than the difficulty of the item. Looking at Figure 1, student *A* represents a student with a low amount of ability since he is only expected to succeed in the three easiest items. Student *B* has an intermediate ability enabling him to succeed in half of the items. Finally, student *C* has the highest amount of ability of the three and is expected to succeed on all items except item *14*, which is the most difficult one. A feature of the scale is that a one unit change in the scale means the same at different parts of the scale.

Figure 1: Subject Ability and Item Difficulty Scale



Source: Rehab-Scales.org.

To put the students on the scale, the latent ability traits are measured, usually by a number of test items, each measuring some facet of the particular ability of interest. Students responding to an item will possess some amount of the underlying ability, which can be considered as a numerical score value.

2.2 Assumptions

IRT is based on a set of assumptions which, if they are met, ensure the validity of the estimates. The two basic assumptions in the IRT models are unidimensional and locally independent item responses. Unidimensionality implies that the number of latent factors equals one. Thus, the set of items only assesses one single underlying trait dimension in accordance to explaining student performance. However, this assumption can never be strictly met because several cognitive, personality and test-taking factors (e.g. motivation, test anxiety, ability to work quickly, tendency to guess when in doubt and so on) always affect test performance. For the unidimensionality assumption to be met adequately the presence of a "dominant" component or factor that influences test performance is required.

The second assumption is local independence. This impose that, conditional on latent factors, items are independent. In other words, if only one ability is assumed to determine success on each item, then the ability level for a given student is the only thing that systematically affect item performance. This implies that when taking all students' abilities into account, there must be no relationship between students' responses to items.

3 Data

The analyses are based on an extraction from the Danish Student Administrative System (STADS). Developed by the Ministry of Education, the Ministry of Research, and a number of companies, STADS is a cross-university system used by all universities in Denmark as well as a range of other educational institutions. STADS is designed to handle anything from admission, enrollment, group formation, registration for courses

and examinations to registration of grades. The database contains detailed information about the study history on a very large number of students and new information is constantly being added. The amount of data is excessive, and there are numerous possibilities of using data.

In this work, administrative data from one cohort of first-year students in the economics program at the University of Copenhagen are used. A total of 262 students were admitted to the program in the spring of 2010. Of these students, 236 participated in at least one exam at the end of the first semester and are therefore a part of the final data.

Two different types of data are considered in this paper. In the ordinal data six different marks are used, ranging on a scale from 0 to 5, where 0 represent the lowest possible value and 5 the highest. Hence, less ability is required of a student to receive the lower marks while more ability is needed to receive the highest. The grades are ordered categorical implying that polytomous model which uses ordinal data contains six different response categories. In binary data the six grade categories from the ordinal data are reduced to two, 0 and 1, where 0 represent the category in which students have the lowest amount of ability. The line dividing the two categories is between the marks 4 and 7 as these are often considered the grades distinguishing the good students from the bad.

The grade −3 is considered a missing observation as there are some problems associated with this lowest possible grade. It is worth emphasizing that −3 is very rarely used, if a student has actually tried to provide an answer to an exam question. Often, the grade -3 is given if a student choses to fail an exam on purpose.

4 Results

This section presents the results of the model estimations. As data are strongly suggestive of a single factor only, unidimensional models are considered in this paper. Initially, the results from the dichotomous two-parameter logistic (2PL) model which uses binary data

is displayed. Following, the results from the polytomous graded response (GR) model, utilizing ordinal data, is shown.

4.1 The Dichotomous Models

The dichotomous 2PL which uses binary data is a simple extension of the Rasch model. Equation (1) is a formulation of the 2PL model which predicts the probability of success $(x = 1)$ for a student as follows:

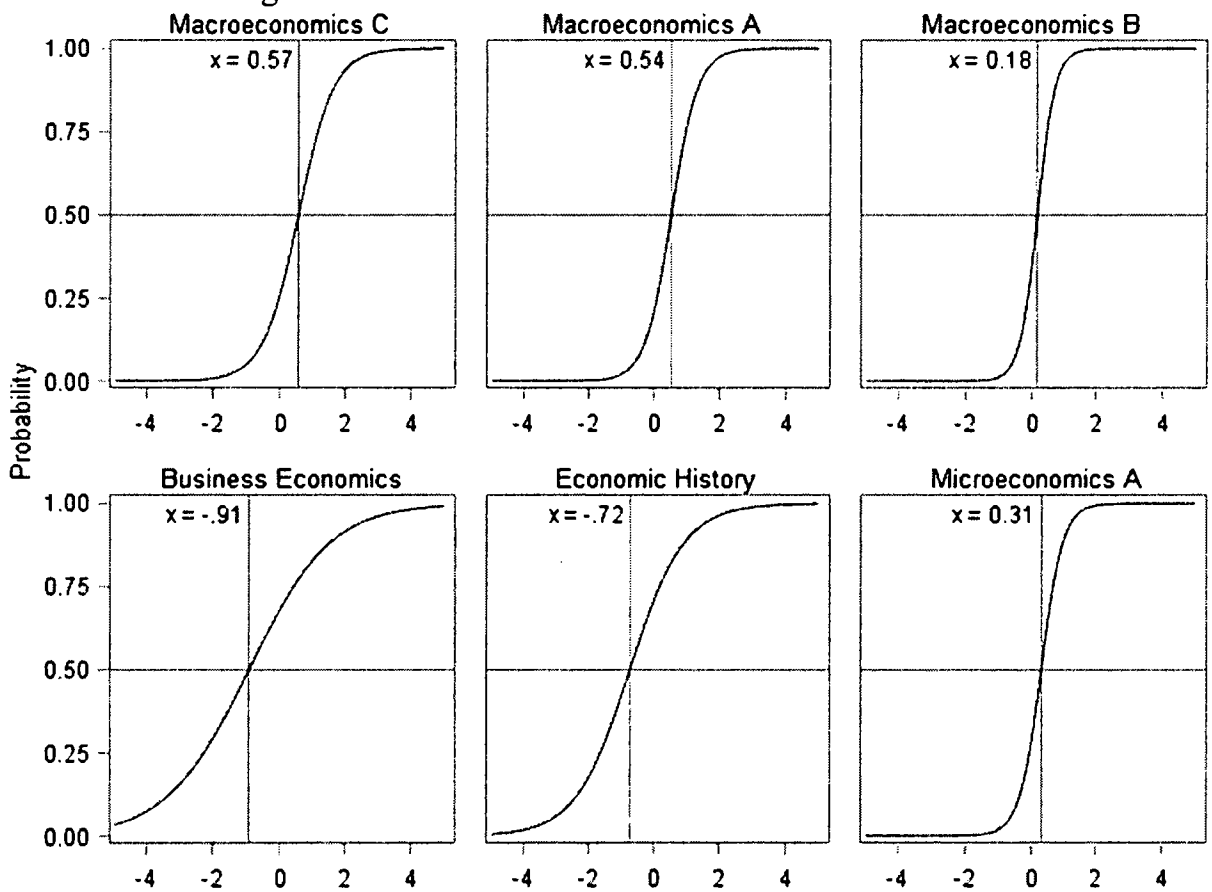
$$P(x = 1|\theta, \alpha, \beta) = \frac{e^{\alpha(\theta-\beta)}}{1 + e^{\alpha(\theta-\beta)}} \quad (1)$$

where θ refers to the ability of the student, β is the difficulty of the item considered and α is the discrimination of the item. Equation (1) further is a representation of the item characteristics curve (ICC), which shows the probability that a student with a given ability will succeed in an item. Generally, students with low ability will have a small probability of succeeding in an item whereas high ability students will have a large probability. Plotting Equation (1) as a function result in smooth S-shaped curves, as the ones depicted in Figure 2. The ICC illustrates the relationship between the probability of succeeding in an item and the ability measured by the test.

The β values from Equation (1) correspond to the point on the ability scale where the ICCs are steepest. This is illustrated in Figure 2 where ICCs for six of the 17 courses are displayed. The ICCs for the two most difficult end-of-course exams, *Macroeconomics C* and *Macroeconomics A*, together with the two easiest courses, *Business Economics* and *Economic History*, from the 2PL model are given in the left side and in the middle of Figure 2. It can be seen, that β also indicates the point on the ability scale, where the probability of getting a grade higher than 4 is 0.5. The α parameter is an estimate of the slope of the ICC. It indicates how good the item is at discriminating between different students. In the right hand side of Figure 2 the ICCs of the two courses with the highest discrimination parameter, *Macroeconomics B* and *Microeconomics A*, are given. The

easiest course, *Business Economics*, additionally has the lowest discrimination indicated by the relatively flat slope of the ICC. It is easier to discriminate between student abilities in *Macroeconomics B* and *Microeconomics A* than *Business Economics*, as the probability of a correct response at the low ability levels are more identical to the high ability levels in *Business Economics.*

Figure 2: Item Characteristic Curves for the 2PL Model



Source: Output from PROC IRT, SAS ® 9.4.

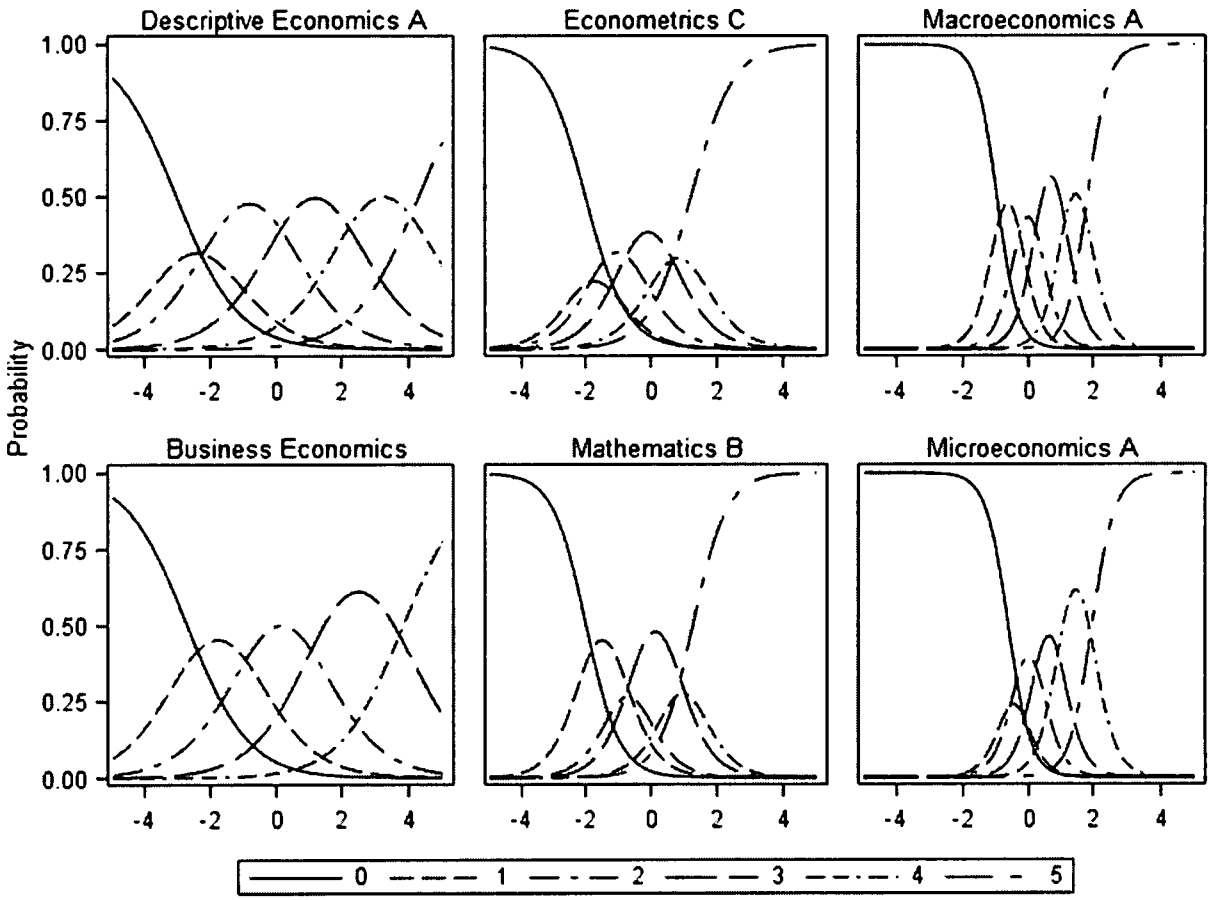
In the dichotomous models the number of categories is limited. However, when imposing a polytomous model, the number of grade categories increases from two to six, leading to an increased amount of information. This might result in an model which is even more precise in estimating ability and has an even better fit of the ability distribution.

4.2 The Polytomous Model

The polytomous GR model is an extension of the 2PL model. Essentially, what occurs in the GR model is that the item is treated as a series dichotomies and 2PL models are estimated for each dichotomy [Embretson et al., 2000, p.99].

The probability that a student with a given ability will score in a particular grade category is illustrated by the category response curves (CRC) given in Figure 3.

Figure 3: Category Response Curves for the GR Model



Source: Output from PROC IRT, SAS ® 9.4.

The two frames at the left side of Figure 3 display the courses where it is most difficult to obtain a high grade, whereas the two frames in the middle are examples of courses where it is relative easy to obtain the highest grade. A difficult course is in this case referring to

the ability levels required to obtain a good grade. The two frames at the right side of Figure 3 illustrate courses with high discrimination, which can be seen because the curves are peaked, making it easier to determine exactly which abilities are associated with each category. These can be compared to the two courses at the left side which, in addition to being the most difficult courses, also are worst at discriminating between ability estimates, indicated by the flat and wide curves.

Figure 3 shows that in *Descriptive Economics A* students need an ability higher than 4 to have equal probability of being assigned the top grade 12 as opposed to any other possible grade. In the opposite end of the grading scale, a very low ability below -3 is needed to have a 50% chance of passing the exam in *Descriptive Economics A*. A surprise appears when considering which end-of-course exams requires the lowest ability to have a 50% chance of receiving the highest grade. When looking at the two middle frames of Figure 3 these courses turn out to be *Econometrics C* followed by *Mathematics B*, where the abilities required are only slightly above 1.

As Figure 3 illustrates, the grade distributions varies quite a lot across courses (CRC for the rest of the courses are available in the author's master's thesis). This indicate, that the grading scales are not used consistently across courses and that the courses varies quite a lot in difficulty. These results provide specific information about which courses are hard for the students, and for which courses the grade distributions are different from the rest. Hence, the university obtain information about which specific courses that might need to be examined in more detail.

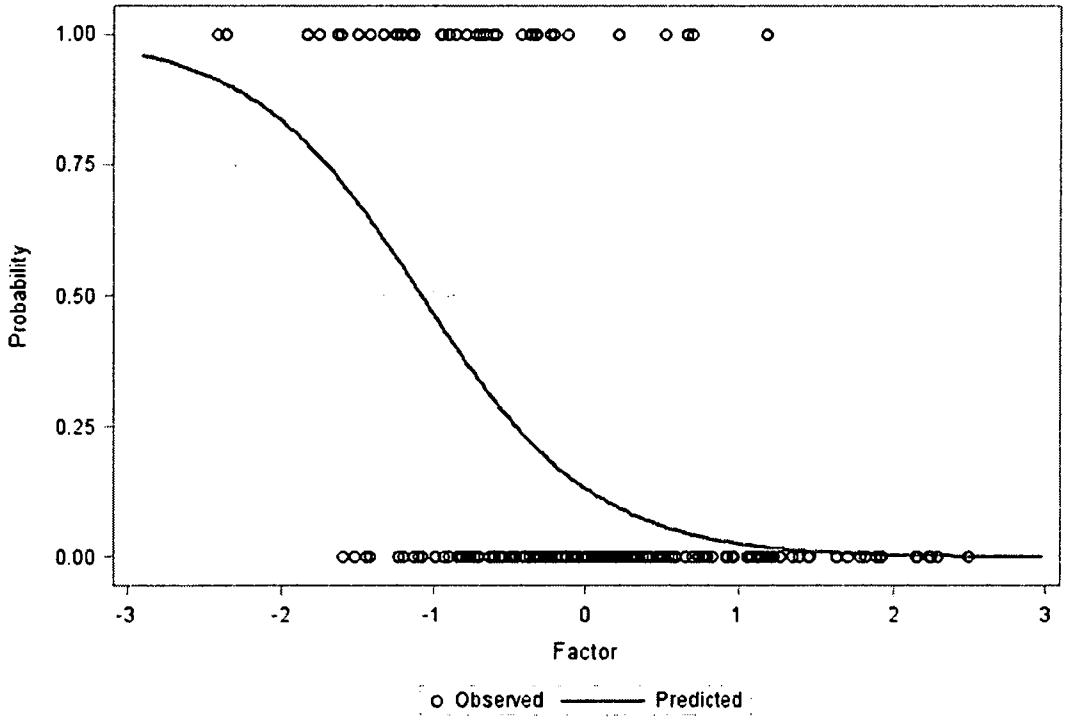
4.3 Education Status

Not all students who undertake an education manage to graduate. In this section the educational status of the students will be examined and measured against the estimated ability parameter from the GR model. This will be done by using a logistic regression model. Initially, the education status for the bachelor's degree is examined followed by a short presentation of the educational status at the master's degree.

4.3.1 Bachelor's Degree Status

To find the amount of student ability required for a student to have a high probability of completing his education, a logistic regression model is estimated. The relationship between the latent ability trait and the bachelor's degree status is illustrated in Figure 4. This figure predicts the probability of a student dropping out of the degree. A clear tendency is seen in Figure 4, showing that students with a lower level of ability have a much larger probability of dropping out of the economics program. Students with an ability level below -1 only have a 50% chance of completing their education. An ability level below -2 results in close to 100% certainty that a student will not manage to graduate. The interesting thing is, as the ability level increases to a value above 1, there is almost no chance that a student interrupts the program. Students with high ability levels who drop out of the program might have experienced a change of interests. They have dropped out from the economics program to attend medical school, for instance.

Figure 4: Predicted Probabilities for Drop Outs at the Bachelor's Degree



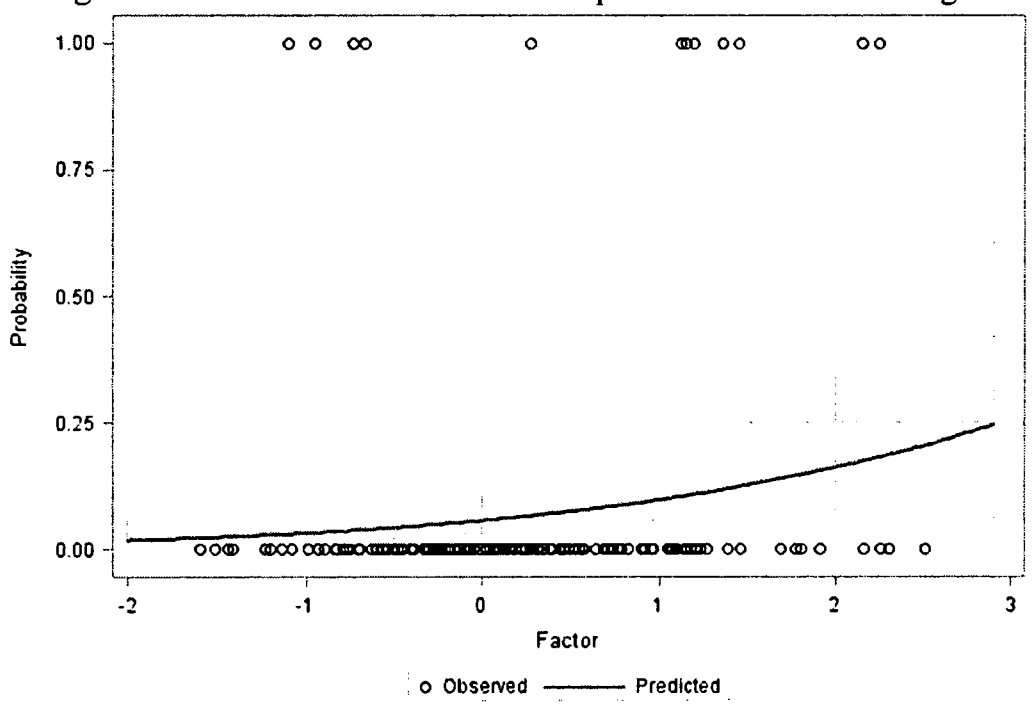
Notes: The circles represent the observed students. The light gray area is the 95% confidence limits.
Source: Output from PROC IRT, SAS ® 9.4.

As can be seen from the confidence limits in Figure 4, the deviations between the predicted model and the observed observation are largest for the low ability levels.

4.3.2 Master's Degree Status

As for the bachelor's degree, a logistic regression is conducted in order to find the relationship between student ability and the probability of dropping out of the master's program. The ability estimates used are the ones obtained from the GR model using all compulsory courses of the bachelor's degree. Figure 5 is a graphical illustration of the relationship. As indicated by the confidence limits, the deviations between the predicted model and the observed observation are considerably larger for the high ability levels. This is mainly caused by the limited number of students, who chooses to interrupt their education. The reason why the high ability students might choose to drop out could be because of a greater number of opportunities. It should be noted that the model has a hard time in predicting the education status for students with a high level of ability.

Figure 5: Predicted Probabilities for Drop Outs at the Master's Degree



Notes: The circles represent the observed students. The light gray area is the 95% confidence limits.
Source: Output from PROC IRT, SAS ® 9.4

4.4 Length of Bachelor's Degree

Applying the ability estimates from the GR model, a linear regression is performed to see how the level of ability relates to the completion time of the bachelor's degree. The calculations are only based on those students who have completed their degree. Table 1 shows the linear regression parameter estimates for the length of the bachelor's degree. When the factor level increases by one, the time a student uses to complete the education decreases by a bit more than a month.

Table 1: Parameter Estimates for Length of Bachelor's Degree

Parameter	*DF*	*Estimate*	*Standard Error*	*t Value*	*Pr > ChiSq*
Intercept	1	3.305	0.029	115.91	<.0001
Factor	1	-0.097	0.020	-4.93	<.0001

Source: Output from PROC IRT, SAS ® 9.4.

5 Conclusion

In this paper, student administrative data on 236 economics students who matriculated at the bachelor's degree at the University of Copenhagen in 2009 have been analyzed to figure out where new initiatives can be implemented in an attempt to improve the educational process for the students. Empirical evidence indicates that the polytomous model outperforms the dichotomous model due to an increased amount of information included in the model.

The approach used in this paper is of course not flawless. Generally it is very difficult to construct a statistical model capable of predicting human behavior. There are however advantages of the approach, as it provides easily interpretable and understandable results, which are simple to communicate to interested parties. All the results (except maybe the result showing that students with high abilities have higher possibility of dropping out of the master's degree) indicate that the ability of students play an important role in the study process.

This paper recommends that the university take closer look at the courses which are troublesome for the students. Further, it is recommended that the university, as soon as possible after matriculation, take care of the weakest students in order to improve the study progress.

References

An and Yung, 2014An, X. & Yung, Y. (2014). *Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It*, Cary, NC: SAS® Institute Inc.

Andersen and Olesen, 2001Andersen, E.B. & Olesen, L.W. (2001). The life of Georg Rasch as a mathematician and as a statistician. In: Boomsma, A., van Duijn, M.A.J., Snijders, T.A.B. (Eds.), *Essays on Item Response Theory*. Springer-Verlag, New York, 3-24.

De Ayala, 2009De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.

Baker, 2001Baker, F. B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD.

Black et al., 2004Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2004). *Working Inside the Black Box: Assessment for Learning in the Classroom*. The Phi Delta Kappan, 86(1), 8-21

Dalskov, 2008Dalskov, M. (2008). *Empirisk Analyse Af Fuldførelsestiden På Politstudiet Ved Københavns Universitet*. Master's thesis.

Embretson and Reise, 2000Embretson, S. E. & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton and Jones, 1993Hambleton, R.K, & Jones, R.W. (1993). *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development*. Instructional Topics in Educational Measurement. (pp.253–262).

Hambleton and Swaminathan, 1985Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers

Hambleton et al., 1991Hambleton, R.K., Swaminathan, H.H. & Rogers, H.J. (1991). *Fundamentals of item response theory* (Measurement methods in the social sciences ; 2). Newbury Park, Calif: Sage Publications.

Henningsen, 2014Henningsen, I. (2014). *Et kritisk blik på PISA*. Presentation at DPU April 2014

Kreiner, 2014Kreiner, S. & Christensen, K.B. (2014). *Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading litteracy*. Psykometrika, vol 79/2:210-231

Mokkink et al., 2010Mokkink, L. B., Knol, D. L., van Nispen, R. A., & Kramer, S. E. (2010). *Improving the Quality and Applicability of the Dutch Scales of the*

Communication Profile for the Hearing Impaired Using Item Response Theory. Journal Of Speech, Language & Hearing Research, 53(3), 556-571.

Rasch, 1960Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests (Studies in mathematical psychology 1)*. København: Danmarks Pædagogiske Institut.

Reckase, 2009Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise et al., 1993Reise, S. P., Widaman, K. F. & Pugh, R. H. (1993). *Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance*. Psychological Bulletin, 114, 552–566.

Schultz, 1961Schultz, T. W. (1961). *Investment in Human Capital*. The American Economic Review 1(2), 1-17

Spearman, 1904Spearman, C. (1904). *"General intelligence," objectively determined and measured*. American Journal of Psychology, 15(2), 201-293.

Templin, 2015Templin, J. (2015). *Item Response Theory - Methods for the Analysis of Discrete Survey Response Data*. ICPSR Summer Workshop at the University of Michigan, from http://jonathantemplin.com (accessed 31.12.2015)

Xu and Jia, 2011Xu, X. & Jia,Y. (2011), *The Sensitivity of Parameter Estimates to the Latent Ability Distribution*. ETS, Princeton, New Jersey. Retrieved October 14, 2015, from http://files.eric.ed.gov/fulltext/ED528984.pdf (accessed 31.12.2015)

Effektevalueringer af sociale indsatser

Mette Foss Andersen, Danmarks Statistik

Indledning

Mit oplæg tager udgangspunkt i en effektevaluering af Christianskolen, som jeg udførte under et praktikophold i Justitsministeriets Forskningskontor fra september 2012 til januar 2013 og som senere dannede grundlag for mit speciale i sociologi. Resultaterne er udgivet i rapporten "Christianskolen -Den kriminalpræventive effekt af et målrettet skoletilbud til unge med særlige behov" (Justitsministeriet 2013c). I denne artikel vil jeg præsentere hvordan det naturlige eksperiment samt matchning er anvendt som grundlag for evalueringen. Jeg vil kort præsentere evalueringens resultater og derefter, med evalueringen som case, diskutere hvilke problemer ved den der kan opstå i forsøget på at dokumentere effekter af sociale indsatser.

Abstract

Følgende er et Mixet methods effektstudie af undervisningstilbuddet Christianskolen. Christianskolen blev etableret som et kommunalt projekt i Frederiksberg Kommune i 2006. Målgruppen er unge i folkeskolens ældste klasser, der har personlige, sociale og/eller psykologiske problemstillinger. Eleverne visiteres til skolen af Pædagogisk Psykologisk Rådgivning (PPR). Omkring 70 procent af de unge har minimum én sigtelse eller mistanke efter straffeloven (kriminalitet begået under den kriminelle lavalder) når de indskrives på skolen.

Skolens vision er at drive "et undervisningstilbud af høj kvalitet der skaber forandring og udvikling". Målsætningen er, at den unge bliver bragt videre i en positiv udvikling personligt, socialt og fagligt, og efterfølgende udsluses til ungdomsuddannelse, praktik, arbejde, jobtræning eller lignende. Desuden er målsætningen et efterværn, der følger eleverne, efter at de er udskrevet fra skolen.

Den kvantitative analyse består af to dele. Den første del er udført som et 'naturligt eksperiment', hvori den registrerede kriminalitet før, under og efter skoleforløbet sammenlignes med en historisk kontrolgruppe. Denne analyse viser, at 27 pct. af de unge fra Christianskolen har begået kriminalitet i en toårig opfølgningsperiode efter skoleforløbet, mens tilsvarende gælder for 54 pct. af de unge i kontrolgruppen. Forskellen i gruppernes kriminalitet er statistisk signifikant ($p = 0{,}033$), når der i en logistisk regressionsmodel kontrolleres for forhold som køn, alder, herkomst og tidligere kriminalitet. En dreng der har gået på Christianskolen har således 37 pct. sandsynlighed for at begå kriminalitet i en to-årige opfølgningsperiode efter

undervisningsforløbet. Tilsvarende gælder for 62 pct. for en dreng i den historiske kontrolgruppe sammenlignet med den historiske kontrolgruppe

I anden del af den kvantitative analyse er der matchet en kontrolgruppe blandt danske lovovertrædere i samme tidsperiode. Kontrolpersonerne er matchet på fødselsår, kvartal, køn og antal tidligere sigtelser. Analysen viser at andelen der begår kriminalitet efter tidspunktet for endt skoleforløb er tilsvarende for både eksperimental- og matchet kontrolgruppe, men antallet af personer med mere end fire sigtelser er højere blandt de matchede kontrolpersoner. Fra kontrolgruppen har 11 personer imellem fem og 27 sigtelser. I eksperimentalgruppen er blot en enkelt person sigtet flere end fire gange.

I evalueringen indgår desuden kvalitative elementer i form af ti enkeltinterview med nuværende og tidligere elever, samt gruppeinterview med henholdsvis lærere, pædagoger, leder og souschef. Desuden indgår deltagerobservation af tre ugers varighed på skolen.

Effektstørrelsen Cohen D peger i begge analyser på en mindre kriminalpræventiv effekt af skoleforløbet og de kvalitative analyser peger i retning af at skoleforløbet kan være kriminalpræventivt. Dog har det ikke været muligt at kontrollere for den generelle faldende ungdomskriminalitet i Danmark i perioden. Desuden er det en plausibel antagelse at matchpersonerne er mere ressourcestærke end eksperimentpersonerne. Dette gør den positive effekt af skoleforløbet statistisk usikker.

Metode og design

Teorien bag eksperimentet

Formålet med effektevalueringer er at udlede sammenhænge mellem en indsats, og et bestemt indsatsmål – i dette tilfælde forløbet på Christianskolen og elevernes efterfølgende kriminalitet.

Ved spørgsmålet om hvorvidt Christianskolen har en effekt på de unges kriminalitet under og efter de afslutter forløbet, er det ikke tilstrækkeligt blot at sammenligne elevernes kriminalitet før og efter skoleforløbet. At elevernes kriminalitet falder efter forløbet, fortæller ikke nødvendigvis noget om en effekt.

Ideelt set skulle kriminalitetsniveauet for hver enkelt elev, der har gået på Christianskolen, sammenlignes med kriminalitetsniveauet for den samme elev, uden at vedkommende havde gået på skolen. Derved ville der blive taget højde for andre faktorer, der påvirker kriminaliteten under forløbet, såsom, alder og miljøet udenfor skolen(Khandker, Koolwal & Samad 2010:22-24). Dette er selvfølgelig umuligt, da et individ ikke simultant kan befinde sig i både det ene og det andet stadie på samme tid. Udfordringen er derfor at finde en kontrolgruppe der er så sammenlignelig med

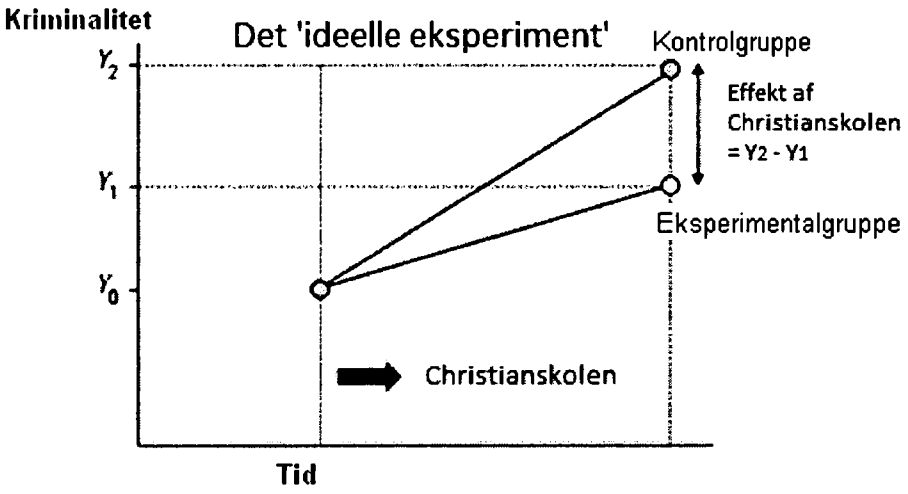
elevgruppen på Christianskolen, at personerne derfra *potentielt kunne* have gået på skolen. Sidstnævnte betegnes også som det kontrafaktiske niveau (Morgan & Winship 2010:5).

Det ideelle eksperiment

I en situation med *det ideelle eksperiment* vil der i interessepopulationen tilfældigt blive udtrukket to forskellige grupper, en eksperimental og kontrolgruppe. Har hver person samme sandsynlighed for at komme i enten eksperimental- eller kontrolgruppe, vil der ikke være systematiske forskelle på de to grupper inden forløbet (Robinson et al. 2009:344; Khandker, Koolwal & Samad 2010:33). Den tilfældige allokering sikrer, at en eventuel funden effekt ikke skyldes andre faktorer end forløbet i sig selv, da den tilfældige udtrækning ikke giver skævhed i grupperne. Det forventes derfor personerne i grupperne har samme udgangspunkt, og udsættes for de samme faktorer der virker udenfor forløbet. Er der et tilstrækkeligt stort antal individer i hver gruppe, kan der antages at være belæg for et kausalt forhold mellem indsatsen og indsatsmålet (Dunning 2008:282; Farrington 2003:162).

Hvis forløbet på Christianskolen kunne fordeles tilfældigt i en eksperimental og kontrolgruppe og antallet af kriminelle handlinger efterfølgende er lavere for eksperimentalgruppen (Y_1) end for kontrolgruppen (Y_2) kan den gennemsnitlige effekt af Christianskolen (δ) for en tilfældig udtrukket person estimeres ved at trække eksperimentalgruppens kriminalitet (Y_1) fra kontrolgruppens (Y_2) (Khandker, Koolwal & Samad 2010:34; Dunning 2008:282).

Figur 1. En situation med det "ideelle eksperiment"[*]



[*]Figur redigeret fra Khandker, Koolwal & Samad (2010:34)

Tilfældigt allokerede eksperimenter er dog sjældne og svære at udføre i praksis, især når det kommer til evalueringer af sociale indsatser. Etiske, praktiske og

økonomiske omstændigheder gør, at det nok aldrig vil være muligt fuldstændigt tilfældigt at opdele en gruppe af unge i to, og tildele den ene gruppe en specifik indsats (Kyvsgaard 2001:11).

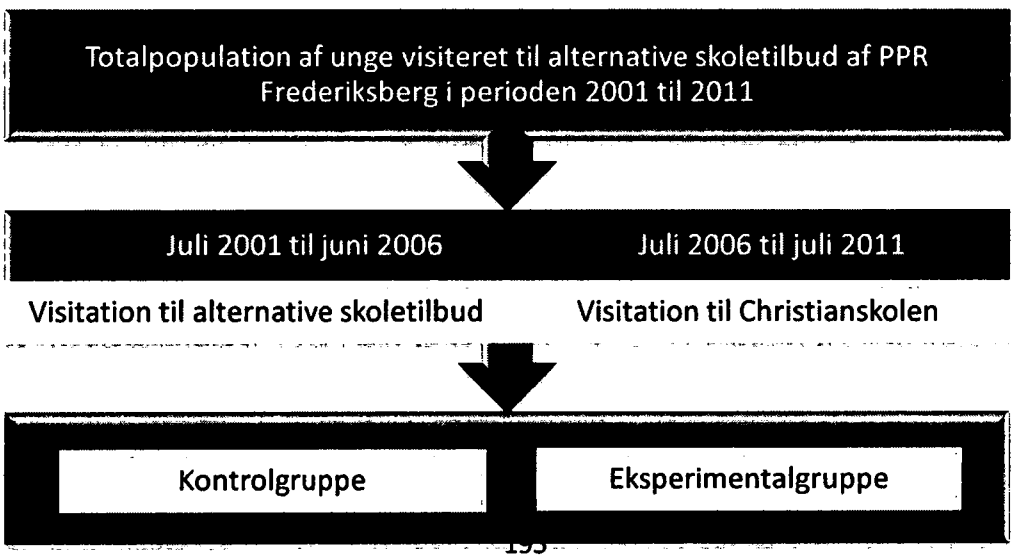
Det "naturlige" eksperiment

Et alternativ til det tilfældigt allokerede eksperiment er det *naturlige eksperiment* (Dunning 2008:291). Til forskel fra eksperimentet fordeles interventionen ved en "naturligt forekommende begivenhed", der mere eller mindre tilfældigt adskiller undersøgelsesgruppen i to tydeligt afgrænsede grupper. Grupper der derefter kan anvendes som eksperimental- og kontrolgruppe (Wooldridge 2010:458; Meyer 1995:151; Robinson et al.2009:346; Dunning 2008:290). I korrekt udførte, naturlige eksperimenter er der belæg for, at tildelingen af interventionen til eksperimental og kontrolgruppe er "as if random", altså "ligesom tilfældigt" (Dunning 2008:283). Dog er denne naturlige, tilfældige tildeling altid mindre perfekt end den fuldstændig tilfældigt allokerede (Robinson et al. 2009:343).

Christianskolen som et "naturligt eksperiment"

Designet til denne undersøgelse er inspireret af det naturlige eksperiment. Christianskolen blev oprettet i 2006 i Frederiksberg Kommune som et undervisningstilbud til unge, der mistrives i de ordinære skoletilbud og har svære problemstillinger. Inden oprettelsen af skolen blev denne gruppe af udsatte unge visiteret til en blanding af offentlige og private skoletilbud både i og udenfor kommunen. Ved oprettelsen af skolen opstod der således en skæringsdato i juli 2006, der opdelte de udsatte unge i to grupper, visiteret før eller efter denne dato. Da det givetvis har været tilfældigt, om hver enkelt ung er blevet visiteret af PPR før eller efter datoen, kan de unge, der forinden blev visiteret til andre skoletilbud, anvendes som en *historisk kontrolgruppe*. De unge, der efterfølgende er visiteret til Christianskolen, anvendes som eksperimentalgruppe.

Figur 2. Forholdet mellem Christianskolen og den historiske kontrolgruppe



Da det er selve oprettelsen af skolen, der fordeler de unge i to grupper, bør der ikke foreligge systematiske forskelle på de to grupper. Kontrolgruppen er udtrukket af PPR, og det er herigennem sikret, at visitationskriterierne har været fuldstændigt ens. De unge fra kontrolgruppen ville ifølge PPR have været visiteret til Christianskolen, i fald visitationen var foregået efterjuli 2006. Og de unge fra Christianskolen ville have været visiteret til andre skoletilbud, i fald visitationen var foregået inden juli 2006. Det er i princippet "ligesom tilfældigt", om en given ung fra den samlede population er blevet visiteret til Christianskolen eller til de alternative skoletilbud.

Kontrolgruppen er i teorien så sammenlignelig med de unge fra Christianskolen, at den tilnærmelsesvist kan opfylde "det kontrafaktiske niveau", altså illustrere hvad der ville være sket med Christianskolens elever, hvis de *ikke* havde gået på Christianskolen.

Ved hjælp af χ^2-teste er der undersøgt om der er signifikante forskelle på grupperne inden indskrivelse på køn, alder, herkomst og antal af tidligere kriminelle forhold. Samlet set viser analyserne af eksperimental- og historisk kontrolgruppe at der ikke foreligger større eller signifikante forskelle. Der er en lille overvægt af piger i eksperimentalgruppen, en forskel der dog ikke er statistisk signifikant.

For at teste om nogle af de førnævnte faktorer alligevel kan have haft indflydelse på visitationen til eksperimental- og kontrolgruppe, er der fortaget en binær logistisk regressionsanalyse af sandsynligheden for at blive indskrevet på Christianskolen.

Tabel 1. Binær logistisk regressionsmodel vedrørende sandsynligheden for at blive indskrevet på Christianskolen. N=89.

Parameter	$\hat{\beta}$	**Standardafvigelse**
Køn	0,781	0,505
Etnisk baggrund	0,172	0,483
Tidligere kriminalitet	0,053	0,051
Alder ved indskrivelse	-0,347	0,219
â	3,815	3,102

note: *** p<0.01, ** p<0.05, * p<0.1

Ingen af modellens parametre har en p-værdi under under 0,1 og PPR's visitation til grupperne kan derfor antages at være tilfældig ud fra de undersøgte parametre.

Matching af alternativ kontrolgruppe

For at understøtte evalueringen er der matchet en kontrolgruppe blandt alle unge lovovertrædere i Danmark i samme tidsperiode. Matchingen er udført blandt data fra Justitsministeriet, der indeholder alle mistanker og sigtelser blandt unge i aldersgruppen 9-17 år fra Danmark i perioden 2001-2013.

Ideen bag matchning et at konstruere et "tvillingepar", hvori den ene person modtager en indsats og den anden ikke gør (Dunning 2008:290). Formålet er at konstruere et kontrafaktisk niveau i form af kontrolpersoner, som ligner eksperimentalpersonerne mest muligt, baseret på en række karakteristika (Khandker, Koolwal & Samad 2010:54).

Matchingen er udført ved hjælp af eksakt matchning og formålet er at finde kontrolpersoner med fuldstændigt identiske værdier på en række variable (Harding, Morgan & Winship 2010:107). De udvalgte matchvariable er følgende: fødselsår, fødselskvartal, køn og antallet af straffelovssigtelser frem til 31. december i det år, den pågældende eksperimentalpersonen er indskrevet på Christianskolen.

I matchningsprocessen er der for hver eksperimentalperson fundet alle mulige kontrolpersoner matchet på de specifikke karakteristika. Herefter er alle de matchede cases bevaret og alle ikke-matchede kontrolcases kasseret. Derefter er der tilfældigt udvalgt tre kontrolpersoner til hver. Kontrolgruppen består dermed af 96 personer.

Data

Datamaterialet er indsamlet via Det Centrale Personregister, Rigspolitiet og Kriminalregisteret i 2013. Eksperimentalgruppen består af alle elever fra Christianskolen, der har afsluttet skoleforløbet fra juli 2006 til og med juli 2011. I alt 52 personer. Den historiske kontrolgruppe er udtrukket af PPR og består af unge, ligeledes fra Frederiksberg kommune, der siden 2001 og frem til skolens opstart i 2006 blev visiteret til andre, daværende skoletilbud. I alt 48 unge. Blandt disse personer måtte i alt fem udgå helt af datamaterialet, da oplysningerne om visitationsdatoer lå for langt tilbage i tiden, og derfor ikke længere var tilgængelige. Yderligere havde seks af de unge havde senere hen også været indskrevet på Christianskolen og derfor allerede indgik i eksperimentalgruppen. Som følge heraf er kontrolgruppen reduceret til i alt 37 personer.

Data til analyser med matchet kontrolgruppe

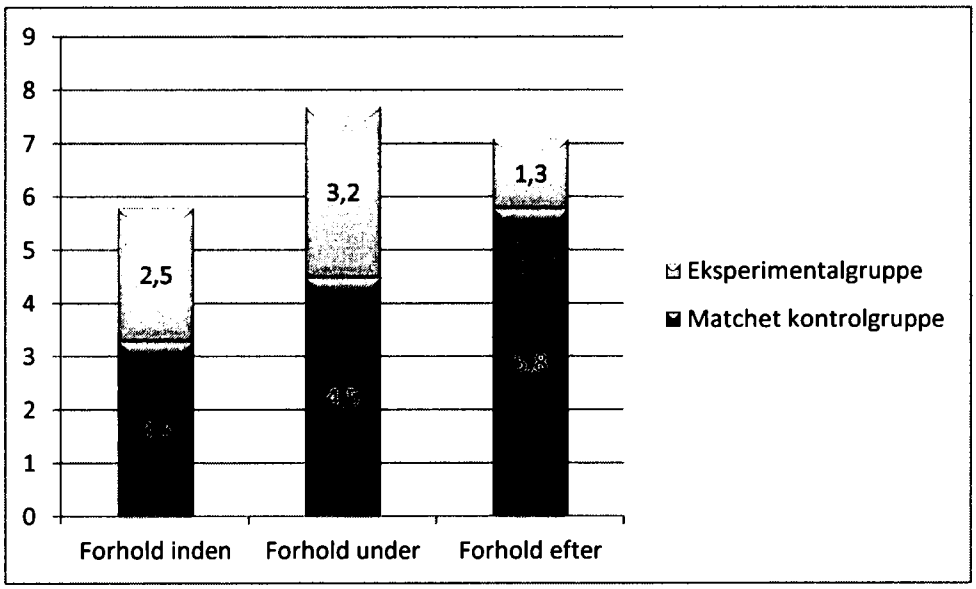
Datamaterialet for den matchede kontrolgruppe udgøres af Justitsministeriets samlede data over omfanget af ungdomskriminalitet i Danmark fra 2001-2013. Eksperimentalgruppen består af de 32 personer fra Christianskolen som havde sigtelser

inden forløbet. Med tre kontrolpersoner per eksperimentalperson består den matchede kontrolgruppe af 96 unge.

Resultater fra den historiske kontrolgruppe

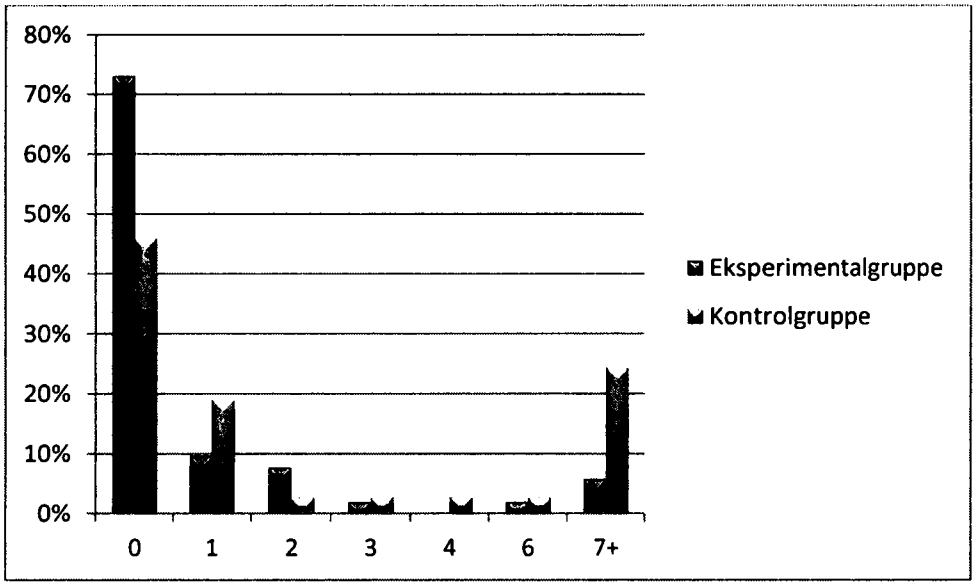
Efter forløbet på Christianskolen falder det gennemsnitlige antal sigtelser pr person. I kontrolgruppen stiger antallet af sigtelser efter de skoleforløb personerne er indskrevet på- Der er i den toårige observationsperiode samlet set registreret 67 forhold efter straffeloven i eksperimentalgruppen og 213 i kontrolgruppen. Kontrolgruppens høje antal sigtelser skyldes blandt andet to personer der tilsammen står for 107 af sigtelserne. Fordelt på grupperne svarer til henholdsvis 1,3 og 5,8 forhold per person, en forskel der er statistisk signifikant (p =0,009). Figur 3 illustrerer det gennemsnitlige antal af sigtelser før, under og efter skoletilbuddene pr. person i eksperimental- og kontrolgruppe.

Figur 3. Det gennemsnitlige antal sigtelser/mistanker per person inden, under og efter interventionen i eksperimental- og historisk kontrolgruppe.



Figur 4 viser den procentvise fordeling af antallet af sigtelser i eksperimental- og historisk kontrolgruppe fordelt efter antal forhold efter skoleforløbet.

Figur 4. Procentvis fordeling af antallet af sigtelser i eksperimental- og historisk kontrolgruppe fordelt efter antal forhold (mistanker/sigtelser) for straffelovsovertrædelser i perioden 2 år efter udskrivning.



Det er især i figurens to ender med henholdsvis 0 forhold og 7+ forhold, at forskellen mellem grupperne viser sig. I eksperimentalgruppen er andelen der ikke er sigtet for kriminalitet, 27 procentpoint større end i kontrolgruppen. Andelen af personer med over syv forhold i perioden udgør 6 pct. i eksperimentalgruppen mod knap en fjerdedel (24 pct.) i kontrolgruppen.

Eftersom at det antages "som tilfældigt" om den unge befinder sig i eksperimental- eller kontrolgruppe, kan et estimat for effekten af at gå på Christianskolen δ_i opnås ved hjælp af modellen (Khandker, Koolwal & Samad 2010:37 Morgan & Winship 2010:129,136):

$$Y_i = \alpha + \delta T_i + \beta X_i + \varepsilon_i$$

$\hat{Y}$ Angiver sandsynligheden for recidiv i opfølgningsperioden. Recidiv defineres ved mindst én sigtelse efter straffeloven i den toårige observationsperiode. Udfald er 1 = Recidiv 0 = Ikke recidiv.

$\hat{\delta}$ Angiver estimatet af effekten af Christianskolen og udfaldet er 1 = Eksperimentalgruppe og 0 = Kontrolgruppe.

Modellens kontrolvariable β består af køn, alder, etnisk oprindelse, længden af indskrivning på skoletilbuddet og tidligere kriminalitet.

Den endelige model er fundet ved hjælp af en forlæns modelsøgning, hvor hvert enkelt parameter inkluderes og fastholdes i modellen, hvis den opfylder kravet om statistisk signifikans. Der er udført modelkontrol. Slutmodellen er angivet nedenfor og indeholder to signifikante parametre: forløbet på Christianskolen og køn:

$$\hat{Y}_{recidiv} = \hat{\alpha} + \hat{\delta} T_{Christianskolen} + \hat{\beta} X_{dreng} + \varepsilon_i$$

Tabel 2 viser resultaterne fra regressionsanalysen.

Tabel 2. Resultat af regressionsanalyse der angiver sandsynligheden for recidiv til kriminalitet efter afsluttet skoleforløb, målt i en toårig observationsperiode. N=89.

	p-værdi	$\hat{\beta}$	Odds	95 % konfidensinterval for
Christianskolen	0,033	-1,021	0,360	[0,141 – 0,918]
Kontrolgruppe			1	
Dreng	0,007	1,543	5,171	[1,566 – 17,080]
Pige			1	
$\hat{\alpha}$	0,061			

I tabellen illustreres p-værdi, parameterestimater og odds-ratioværdier. Odds-ratioværdien på 0,36 for parameteret Christianskolen betyder således, at de unge, der har gennemgået forløbet på Christianskolen, har 64 pct. mindre sandsynlighed for at begå kriminalitet end de unge i kontrolgruppen. Konfidensintervallet spænder dog bredt imellem værdierne 0,141 og 0,918. Det viser, at den estimerede effekt både kan være relativt lille (8,2 pct.) og meget stor (86 pct.).

Parameterestimaterne, $\hat{\beta}$, kan bruges til at udregne sandsynligheder i grupperne. I tabel 14 illustreres sandsynlighederne for de to grupper:

Tabel 3. Sandsynlighed for Y = 1 i opfølgningsperioden

	Eksperimentalgruppe	Kontrolgruppe
Dreng	0,37	0,62
Pige	0,10	0,24

Tabellen viser, at en dreng fra Christianskolen har 37 pct. sandsynlighed for at recidivere i opfølgningsperioden. Tilsvarende er sandsynligheden blandt en dreng i kontrolgruppen 62 pct.

Et andet effektmål er Cohens D, som er et standardiseret effektmål beregnet til at sammenligne effekter på tværs af evalueringer. Dog skal effektmålet tages med visse forbehold, blandt andet fordi, der kun måles på årsag, δ, og effekt, Y, og ikke tages højde for kontrolvariable (Schuele & Justice 2006). Effektmålet, $\hat{d}$, repræsenterer gennemsnittet skaleret af standardafvigelsen i den undersøgte population (Howell 2011:366). Tolkningen af $\hat{d}$ er illustreret i tabel 3 nedenfor (Schuele & Justice 2006).

Tabel 4. Tolkning af Cohens D

$\hat{d}$	Effektstørrelse
0,2	lille effekt
0,5	medium effekt

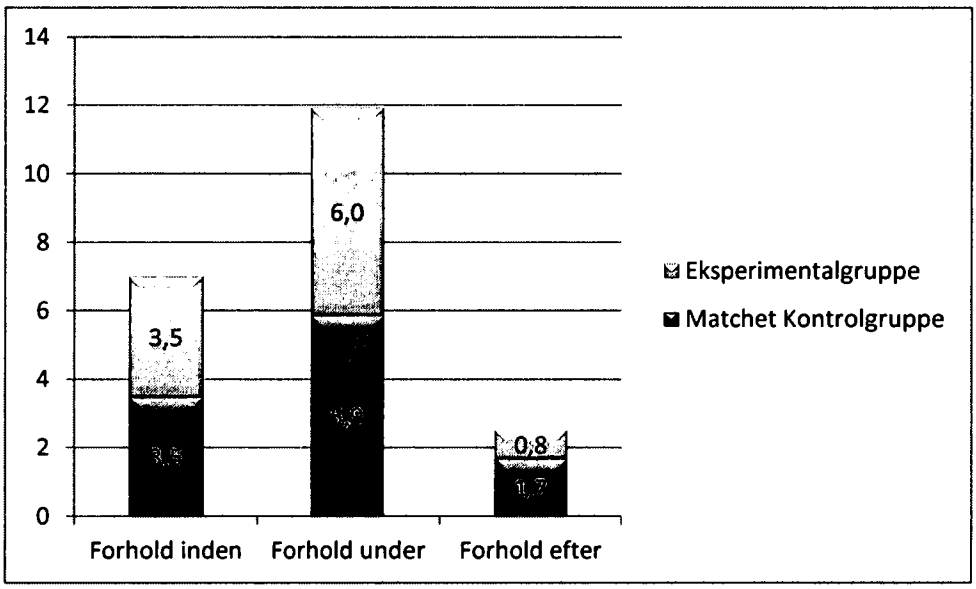
0,8	stor effekt

I analysen af den historiske kontrolgruppe er $\hat{d}$=0,47, et sted mellem "lille" og "medium" effekt.

Resultater fra matchet kontrolgruppe

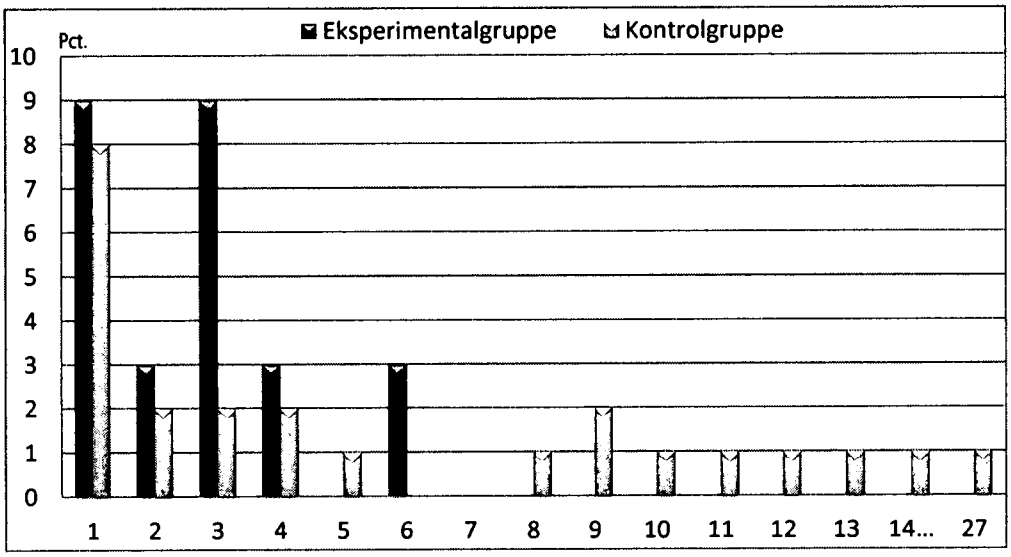
Figur 5 illustrerer de gennemsnitlige antal forhold før, under og efter indskrivnings- og udskrivelsesdatoerne for henholdsvis eksperimental- og matchet kontrolgruppe.

Figur 5. Det gennemsnitlige antal sigtelser/mistanker per person inden, under og efter interventionen i eksperimental- og matchet kontrolgruppe.



I denne analyse deltager kun elever fra Christianskolen der havde sigtelser/mistanker/mistanker inden de påbegyndte forløbet. Blandt disse er i alt ni personer tilsammen sigtet for 24 forhold efter udskrivelse, mens 26 personer fra kontrolgruppen tilsammen er sigtet for 160 forhold. Dette svarer til 0,8 forhold per person i eksperimentalgruppen og 1,7 i kontrolgruppen - en forskel der ikke er statistisk signifikant. Efter interventionen har personerne i den matchede kontrolgruppe i gennemsnit knap 1 forhold mere i opfølgningsperioden.

Figur 6. Den procentvise fordeling af mistanker og sigtelser i eksperimental –og matchet kontrolgruppe i den toårige opfølgningsperiode.



Figur 6 illustrer den procentvise andel af sigtelser for blandt de 9 personer fra Christianskolen og 26 personer fra kontrolgruppen der er sigtet i opfølgningsperioden. Som ved den historiske kontrolgruppe, er andelen med et højt antal af sigtelser lavere i eksperimentalgruppen end kontrolgruppen. I alt 11 personer fra kontrolgruppen har imellem 5 og 27 sigtelser. En enkelt person fra eksperimentalgruppen er sigtet mere end fire gange.

En udregning af Cohens D giver værdien 0,25, hvilket peger på en "lille" effektstørrelse.

Niveauerne i "Scientific Methods" skalaen

Henriette Christensen fra det Kriminalpræventive Råd, har kortlagt effekter af kriminalpræventive indsatser ud fra "Scientific Methods"-skalaen (Christensen 2012:45). Skalaen er udviklet af Sherman (et al. 1998), og i denne placeres undersøgelser på niveau 1-5 i en vurdering af, hvor høj intern validitet studiet har. Intern validitet angiver sikkerheden for, hvorvidt en effekt kan tilbageføres til den evaluerede indsats eller program (Sherman et. al 1998:3; Christensen 2012:45). Med inspiration fra Christensen (2012:46) har jeg oversat skalaen fra Sherman (et al. 1998:4-5), illustreret i tabel 5.

Tabel 5. Opsummering af "Scientific methods"-skalaens 5 niveauer

	Kriterie i Scientific Methods skalaen
1	Indsatsmål (Y) efter indsats (X)
2	Indsatsmål (Y) inden og efter indsats (X)

3	Indsatsmål (Y) inden og efter indsats (X) og kontrolgruppe
4	Indsatsmål (Y) inden og efter indsats (X), kontrolgruppe, og kontrolvariable som antages at påvirke effekten (Z)
5	Indsatsmål (Y) inden og efter indsats (X) på tilfældigt udvalgt eksperimental- og kontrolgruppe

Niveau 1 repræsenterer det studie, der antages at have lavest intern validitet og niveau 5 det højeste - det tilfældigt allokerede eksperiment. Undersøgelsen af Christianskolen kan placeres på niveau 4, da der er målinger inden og efter indsatsen, sammenligning med kontrolgruppe, og der kontrolleres for faktorer, der i andre studier har vist sammenhæng med kriminalitet. Med inspiration fra Scientific Methods-skalaens niveauer vil jeg i det følgende diskutere og vurdere undersøgelsens interne validitet.

Antal observationer i den kvantitative analyse

Styrken af en effektevaluerings interne validitet afhænger af hvor godt et kontrafaktisk niveau kontrolgruppen er. Som nævnt tidligere antages kontrolgruppen ideelt set som at være et 'spejlbillede' af eksperimentalgruppen. Begivenheder under forløbet, som påvirker kriminalitetsniveauet, antages i teorien at ske parallelt for begge grupper. Men spørgsmålet er, om denne antagelse holder, når undersøgelsesgrupperne ikke er større. Et belæg for et kausalt forhold kræver netop, at der er tilstrækkeligt med individer i hver gruppe (Dunning 2008:282; Farrington 2003:162). Det er især væsentligt, at alle personer i de to grupper er lige disponerede for at begå kriminalitet. Hvis der i den ene af grupperne er en overvægt af unge, der på forhånd ikke er disponerede til at begå kriminalitet, vil individerne i denne gruppe sandsynligvis også være mere modtagelige overfor det pædagogiske arbejde. Disse unge vil sandsynligvis også være mere motiverede til at fralægge sig en eventuel kriminel adfærd (Khandker, Koolwal& Samad 2010:26;Morgan&Winship 2010:40,47).

Selektionsbias ved den matchede kontrolgruppe

Styrken ved matchning er, at der fra en udvalgt undersøgelsespopulation kan konstrueres en kontrolgruppe ud fra nogle veldefinerede karakteristika. Udfordringen er at inkludere variable der understøtter det kontrafaktiske niveau. For selvom to personer er født i samme kvartal, har samme køn, og har haft samme antal sigtelser på samme dato, har de ikke nødvendigvis samme vilkår.

I undersøgelsen har det ikke været muligt at matche på socioøkonomiske faktorer, da der ikke har været de tilgængelige data. Dette kan give selektionsproblemer i analysen, da der ikke er sikkerhed om, hvorvidt den matchede kontrolgruppe har haft samme socioøkonomiske vilkår. En sammenligning af den kriminalitet som henholdsvis eksperimental- og kontrolgruppe har begået inden forløbet på

Christianskolen, viser at kriminaliteten blandt Christianskolens elever er af noget grovere karakter end kriminaliteten blandt de tilfældigt udvalgte matchpersoner i resten af Danmark. Eksempelvis er eksperimentalgruppen i højere grad registreret for indbrud og simpel vold, og i kontrolgruppen i højere grad er sigtet for mere generel ungdomskriminalitet som butikstyveri og hærværk. Forskellen i kriminalitetens art skyldes sandsynligvis at alle elever fra Christianskolen har været visiteret af PPR, der typisk beskæftiger sig med børn og unge med særlige behov (Undervisningsministeriet 2000). For personerne i den matchede kontrolgruppe er der derimod ingen information om, hvorvidt de også har været i forbindelse med PPR eller har samme problemstillinger som de unge på Christianskolen. Det er dermed en plausibel antagelse at kontrolgruppen ikke er lige så udsat som eleverne fra Christianskolen og at risikoen for kriminel adfærd i eksperimentalgruppen på forhånd også er større.

Den amerikanske forsker Robert Bifulco (2012:731,749) peger på, at selektionsbias kan opstå grundet "geografisk mismatch", hvis eksperimental - og matchet kontrolgruppe kommer fra områder, der er demografisk heterogene. Selvom Danmark ikke på samme vis som USA har områder med store skel imellem indkomst- og uddannelsesniveauer, findes der regionale forskelle når hovedstaden sammenlignes med resten af landet. Eksempelvis er antallet af sigtelser per indbygger det højeste i København (Justitsministeriet 2013b:6), ligesom der i København og omegn er og har været et langt større omfang af visitationszoner. Faktorer, der muligvis kan have betydning for antallet af sigtelser og mistanker i henholdsvis eksperimental- og kontrolgruppe.

Konsekvenserne af forskelle i de to gruppers selektion kan betyde at en eventuel effekt af Christianskolen fremgår mindre tydelig. Havde det været kontrolgruppen, der antageligvis var i større risiko for kriminel adfærd, ville det derimod være et problem for resultaterne, da en funden effekt ville synes upålidelig.

Det vil i matchning altid være et vilkår, at der findes betydningsfulde kvantitative og kvalitative faktorer som ikke kan inkluderes i selektionen og som påvirker resultaterne i større eller mindre grad (Khandker, Koolwal & Samad 2010:54).

Periodeeffekter ved historiske kontrolgrupper

Et typisk problem for historiske kontrolgrupper er periodeeffekter i form af hændelser, der foregår under forløbet, og som kan påvirke effekten i observationsperioden. Periodeeffekter kan medføre, at eksperimental- og kontrolgruppen ikke har opereret under samme vilkår, og at der foreligger alternative forklaringer på resultaterne (Meyer 1995:152). I denne undersøgelse kan der være en periodeeffekt, da ungdomskriminaliteten i Danmark generelt er faldet med 28 pct. fra 2001-2012 (Justitsministeriet 2013:7). Det generelle fald i kriminaliteten kan betyde, at det reducerede kriminalitetsniveau blandt Christianskolens elever ikke alene kan isoleres til forløbet på skolen.

Det har ikke været muligt at kontrollere for den faldende ungdomskriminalitet i studiet. Hvis en del af forklaringen på den faldende kriminalitet i eksperimentalgruppen skyldes et generelt fald i ungdomskriminaliteten, lider parameterestimatet af upward bias (Gujarati & Porter 2010:222). Parameterestimatet $\hat{\delta}_{\text{Christianskolen}}$, vil således være biased og højere end den "sande" effekt af $\delta_{\text{Christianskolen}}$.

En anden potentiel periodeeffekt er politi- og retsreformen fra 2005, som blandt andet har betydet mere politi på gaden, ligesom der siden 2004 har været visitationszoner i København og omegn. Det kan betyde, at eksperimentalgruppen har haft større sandsynlighed for at modtage sigtelser end kontrolgruppen ville have haft, hvis de havde været "aktive" i samme tidsperiode. Dette kan føre til downward bias på parameterestimatet $\hat{\delta}_{\text{Christianskolen}}$(ibid.). Estimatet $\hat{\delta}_{\textit{Christianskolen}}$ ved således være er lavere end den "sande" effekt af $\delta_{\text{Christianskolen}}$.

Uobserverede faktorer i effektstudier

I studier hvor eksperimental og kontrolgruppe ikke er perfekt tilfældigt allokerede vil det altid være et problem at eliminere alternative forklaringer (Farrington 2003:163). Spørgsmålet er om det så overhovedet er muligt at bedrive kausal videnskab? Ifølge Robinson (2009:344) er det muligt, hvis den kausale mekanisme virker plausibel. Til netop at synliggøre kausale forhold mellem variable er triangulering af metoder et godt arbejdsredskab (Christensen 2012:50). Ved hjælp af de kvalitative data har jeg kunne identificere eventuelle kausale mekanismer mellem forløbet på Christianskolen og de unges kriminalitet, som ikke har kunne analyseres ud fra de kvantitative data. Ifølge Christensen (2012:51) har effektstudier oftest enten øje for effekt eller for proces, men netop ved en undersøgelse af selve indsatsen og effekten kan de to faktorer sættes i relation til hinanden. Analysen af de kvalitative data, viste også at de forskellige tiltag, strategier og metoder Christianskolen anvender i arbejdet med eleverne, i andre undersøgelser har påvist en kriminalpræventiv effekt. Er der teoretiske belæg for at indsatsen påvirker indsatsmålet kan det ifølge Robinson (2009:344) anvendes som belæg for at de to faktorer er meningsfuldsfuldt relateret.

Litteratur

Christensen, Henriette Nobili 2012: *Effekten af Mentor- og fritidsindsatser for unge i risiko. En systematisk kortlægning.*Det Kriminalpræventive Råd.

Dunning, T.2009: "Improving Causal Inference: Strengths and Limitations of Natural Experiments".*Political Research Quarterly*, 61: 282-293

Farrington, David P. 2003 : British Randomized Experiments on Crime and Justice. *Annals of the American Academy of Political and Social Science*, vol 589: 150-167.

Frederiksberg Kommune 2008: *Virksomhedsplan Pædagogisk Psykologisk Rådgivning, Frederiksberg*. Frederiksberg Kommune.

Harding, David, Morgan, Stephen L. &Winship, Christopher 2010 "Matching Estimators of Causal Effects" in *Counterfactuals and Causal Inference: Methods and Principles for Social Research.*Cambridge University Press.

Howell, David C. 2011: *Fundamental statistics for the behavioral sciences*. 7. International ed.Cengage Learning

Justistministeriet 2013:*Udviklingen i børne- og ungdomskriminalitet 2001-2012. Med separate opgørelser for kommuner og politikredse.*Justitsministeriets forskningskontor.

Justistministeriet 2013b: *Kriminalitetsniveauet i kommuner og politikredse 2011.* Justitsministeriets Forskningskontor.

Justitsministeriet 2013c: Christianskolen. Den kriminalpræventive effekt af et målrettet skoletilbud til unge med særlige behov.Justitsministeriets Forskningskontor.

Khandker, Shahidur R.; Koolwal, Gayatri B. &Samad, Hussain A. 2010: *Handbook on Impact Evaluation.Quantitative Methods and Practices.*The World Bank Washington.

Kyvsgaard, Britta 2001: "Bliver de mindre kriminelle af det? –effekten af tiltag over for lovovertrædere". *Social politik* 3. 1

Meyer, Bruce D. 1995: "Natural and quasi-experiments in economics" Journal of Business & Economic Statistics; April 1995; 13, 2, . p. 151

Robinson, G.; McNulty, J.E. og Krasno, J.S. 2009 : "Observing the Counterfactual? The Search for Political Experiments in Nature", Political Analysis, 17: 341–357

Schuele, C.M & Justice, L. M 2006: The Importance of Effect Sizes in the Interpretation of Reseach: Primer on Research: Part 3. The ASHA Leader.

Sherman, *Lawrence W. , Denise C. Gottfredson, Doris L. MacKenzie, John Eck, Peter Reuter, and Shawn D. Bushway* 1998: Preventing Crime: What Works, What Doesn't, What's Promising. *A report to the United States Congress*, Department of Criminology and Criminal Justice,University of Maryland.

Should I stay or should I go –

Hvorfor forlader danske kystfiskere erhvervet?

Ayoe Hoff, Rasmus Nielsen, Max Nielsen
Institut for Fødevare- og Ressourceøkonomi
Københavns Universitet

Det danske kystfiskeri er som den resterende del af dansk fiskeri i tilbagegang. Årsagen er den nye regulering med individuelt omsættelige kvoter indført i 2003-2007. Det på trods af at der fra politisk hold er gjort meget for netop at støtte op om små fiskefartøjer (mindre end 17 meter lange) der fisker tæt ved deres hjemhørshavn. Der tildeles ekstra kvoter til små fartøjer for hvilke 80 % af deres fangstrejser er på max 3 døgn, og disse øremærkede kvoter forøges hvis fartøjerne samtidig fisker med bæredygtige redskaber. Alligevel er andelen af fiskere tilknyttet små fartøjer faldet med mere end 50 % i perioden 2002 til 2012. I dette studie undersøges ved hjælp af logistisk regression årsagerne til at danske kystfiskere forlader erhvervet. Løn og socioøkonomiske data fra alle personer der på et tidspunkt i perioden 2002-2012 har været beskæftiget i kystfiskeriet inddrages, og beslutningen om at forlade fiskeriet holdes op mod bl.a. personens alder, om han er lønmodtager eller selvstændig, indtægt før og efter fiskeriet forlades, sociale ydelser og pension før og efter fiskeriet forlades, familie indkomst, plus type af fartøj fiskeriet foregår på.

Public Procurement in the EU. Another explorative study.

by
Lisbeth la Cour, Department of Economics, CBS
and
Grith Skovgaard Ølykke, Department of Economics, CBS

1. Introduction.

As often stated in publications that relates to public procurement in the EU this field of study is of major importance as we are talking 'big money': a percentage of GDP around 16, Commission (2011). Data that relates to this field can be found in Tenders Electronic Daily provided by the Commission and compared to earlier studies of these data, see e.g. la Cour & Milhøj (2013) or la Cour, Milhøj and Ølykke (2015), the access to the data has become somewhat easier over the last couple of years.

The idea of the public procurement directives is to increase the competition to increase efficiency and welfare, Commission (2011), by enlarging the market when a tender has a sufficient large size. The exact size depends on the type of the tender. Below we will describe some features that concerns competition based on the measures made available to researchers through Tenders Electronic Daily[1].

There is no perfect measure of competition available in Tenders Electronic Daily but in most cases we know the number of bids for a contract and even though we are aware that such a number at most can be considered necessary but not sufficient to judge the degree of competition in a field, we believe that valuable knowledge can be derived by studying the number of bids. Or as we have chosen to do in the present paper: to study cases characterized by a lack of competition when measured by the number of bids for a contract.

The paper will continue with some more information on the data base in section 2. In section 3 we start our descriptive analysis by provide the broad

[1] We would like to thank the Commission for making the data available to us.

picture. In section 4 we try to dig a little deeper and finally, section 5 contains some tentative conclusions.

2. The data

As mentioned above we use the data from Tenders Electronic Daily (TED) in the present study. TED contains many different types of text documents from, Contract Notices (CN), Contract Awards (CA) over transparency notices to Contract Award Notices (CAN). Here we will focus on the second and the last types, the contract awards and the contract award notices. In their original form the CANs are text documents but information collected electronically is now available from two different sources: The EU commission and also a group of researchers who have constructed a data base called OpenTED (See the WEB-page reference in the list of references). Each CAN may have more than one award and in both of the electronic data bases each record will identify a single award and hence each CAN may be present more than once in the data set. Many of the variables in the two data bases are the same but there are exceptions. Only the data from the EU commission contains an identifier variable for a framework agreement and for dynamic purchasing systems. Also the data from the commission includes information that goes back to 2009 while the OpenTED starts in 2012 – actually at the end of January 2012. Therefore one could argue for just applying the data from the commission but we have discovered some problems with a couple of CANs missing around the turn of each year so in the end we decided to use the OpenTED as our main source of information. Due to 2015 at the point of writing still being incomplete we decided to stop our sample at the end of 2014. The next section will show some initial figures on the number of CANs and CA's.

Our data are delimited to contain CANs from all EU and EEA countries hence Norway, Liechtenstein and Iceland are added to the sample of the 28 EU countries. Switzerland is not an EEA country and therefore not part of our sample. The kind of variables we have access to are: identity and type of the

procuring authority, identity of the winners, the number of bids, the directive behind the CAN, the type of work, the type of procedure, the award criteria, the location of the unit of the CAN and the industry of the CAN classified by the EU CPV codes. Also a broader classification of type of field can be found. We do not have information on all these variables for all CANs.

Table 1: The number of CAN's and CA's for our sample period.

Year	2012	2013	2014
CAN's	157499	160356	163245
CAN in OpenTED data	141303	159521	161272
CANs in our study: (Only EU and EEA countries and in OpenTED)	139946	157806	159237
CAs in our study	430665	482358	503589
CA's with missing number of bids	78300 (18.2%)	88070 (18.3%)	93171 (18.5%)
CA's with number of bids equal to zero	1366 (0.3%)	1073 (0.2%)	600 (0.1%)
CA's with numbers of bids equal to 1	88097 (20.5%)	98681 (20.5%)	103430 (20.5%)
CA's with number of bids larger than 1	262902 (61.1%)	294534 (61.1%)	306388 (60.8%)

Note: Remember that not all of 2012 are covered in OpenTed. OpenTed starts on 31 January 2012.

3. The broad picture.

In table 1, we will start by providing the number of CANs for the years 2012, 2013 and 2014. These numbers are extracted directly from the on-line version of TED.

In OpenTED only documents that obey the standard structure of the documents with well-defined sections are included. In online TED there are some documents that do not obey to this structure (mainly related to EuropeAid – but not exclusively). We restrict our study to be based on the

documents in OpenTED only. Also we focus on documents for EU and EEA countries. Finally it is worth noticing that OpenTED contains no documents with a publication date prior to 31 January 2012. The first document in OpenTED has the number 31450 and belongs to 31 January 2012. We have not tried to include documents with a number smaller than 31450. The total number of CANs over the sample period is 481100 and the total number of CAs over the sample period is 1416612. It is the CAs that will provide our unit of analysis as we have information of number of bids and hence some indication of the level of competition at this level. The overall impression from table 1 is a small increase in the number of CANs. Also the number of CAs increases over the years.

Next we turn our interest toward the information that concerns the competition for each contract award. The European Commission has chosen to focus on the average number of bids per contract and states that in general the competition seems satisfactory from the point of view of the Commission if the average is high. They observe, however, that in almost 20% of the cases only 1 bid was noticed, see Commission (2011). It is, however, worth noticing that the share of CAs with just 1 bid and the share of CAs with more than 1 bid seem to be very stable. Also it is worth noticing that the share of CAs with just one bid is fairly high i.e. around 20% in each of the years.

When taking a closer look at the CAs that signals no competition we first observe that only a limited number of CAs have 0 bids. We will disregard these all together later in our study as this must imply that no winner was found. At the moment we keep them as a separate category in the tables.

The second thing to be mentioned before digging a bit deeper into the data concerns the distribution of bid types over the countries. For this purpose we collect all the years into a single sample and choose some of the countries that show patterns of interest to be displayed in table 2. There are large differences amongst the countries and the purpose of table 2 is to give some indication of this without going into too many details.

Table 2: Distribution of number of bid categories by country.

Country	Missing	0 bids	1 bid	More than 1 bid	Total
France (FR)	38.3%	0.2%	9.6%	52.0%	351916
Poland (PL)	3.5%	0.0%	44.4%	52.1%	358727
Germany (DE)	17.2%	0.4%	10.1%	72.3%	102629
UK	18.0%	0.1%	4.6%	77.4%	94307
Italy (IT)	30.4%	0.7%	19.2%	49.7%	57253
Spain (ES)	42.4%	0.2%	10.7%	46.8%	52807
Romania (RO)	0.1%	0.0%	18.8%	81.1%	61842
Bulgaria (BG)	1.8%	0.0%	23.4%	74.8%	32069
Denmark (DK)	29.4%	0.0%	5.7%	65.0%	15648
Total - also countries not in the table	18.3%	0.2%	20.5%	61.0%	1416612

As in total we have 31 countries in our sample we have chosen not to include all countries in the table. The countries selected for table 2 are the larger countries in terms of number of CA's and Denmark. If focusing on the column showing the percentage for a country in case of just 1 bid we see quite large differences. In case of 1 bid there seem to be groups of countries around different levels. Countries like Denmark, Germany, Spain, Finland, France, Ireland, Iceland, Liechtenstein, Luxembourg, the Netherlands, Norway, Portugal, Sweden and the UK have percentages around 10 and below, while Austria, Belgium, Czech Republic, Estonia, Greece, Italy, Lithuania, Latvia, Malta, Romania, Slovenia and Slovakia have between 13 and 16%. A few countries stand out from the rest with very high percentages of contracts with just 1 bid: Croatia, Hungary and especially Poland. For Poland the combination of a very high 1-bid percentage of 44.4 and a very large number of contract awards points towards a case that deserves more attention later on.

Next we turn to the directives that have been used for the document. In table 3 we show the distribution of classes of bids across directives.

Table 3: Distribution of bids by directive.

	missing	0	1	more than 1	total
2004/17	24.1%	0.1%	18.3%	57.6%	44719
2004/18	18.1%	0.2%	20.5%	61.1%	1367665
2009/81	20.7%	0.2%	27.9%	51.2%	4228
Total	259541	3039	290208	863824	1416612

Note: 2004/17 is the service utility directive; 2004/18 is the public purchase directive and 2009/81 is the defense directive.

A test for similar distributions clearly reject this hypothesis supporting the visual impression from the table.

A cross tabulation of bid distribution against the type of contract shows the results provided in table 4:

Table 4: Distribution of bids by contract type.

	missing	0	1	more than 1	total
Services	18.4%	0.2%	18.7%	62.7%	489137
Supplies	17.0%	0.2%	24.3%	58.5%	781512
Works	25.4%	0.3%	5.8%	68.6%	145963
Total	259541	3039	290208	863824	1416612

The general picture from table 4 is that the one bid case is most prevalent for the Supplies.

A cross tabulation of bid distribution against contract authority type is provided in table 5. The highest percentages of more than 1 bid is seen for 'National or Federal Agency/Office' and the 'European Institution/Agency or International Organisation' category while the highest percentages for 1-bid cases are seen for 'Body governed by public law' and 'other'.

Table 5: Distribution of bids by contract authority type.

	missing	0	1	more than 1	total
Ministry or any other national or federal authority	14.0%	0.3%	20.9%	64.8%	119505
Regional or local authority	19.8%	0.2%	14.1%	65.9%	310585
Utilities	24.0%	0.1%	18.3%	57.6%	44774
European Institution/Agency or International Organisation	5.8%	0.0%	18.5%	75.7%	6612
Body governed by public law	12.4%	0.2%	26.0%	61.4%	447210
Other	14.8%	0.1%	26.2%	58.9%	294788
National or federal Agency/Office	9.1%	0.5%	11.4%	79.1%	24890
Regional or local Agency/Office	25.0%	0.3%	12.4%	62.3%	30449
Not specified	44.2%	0.6%	4.1%	7.4%	137799
Total	259541	3039	290208	863824	1416612

Table 6: Distribution of bids by procedure.

	missing	0	1	more than 1	total
Open procedure	17.4%	0.2%	20.4%	62.0%	1212264
Restricted procedure	17.3%	0.1%	6.8%	75.8%	64859
Accelerated restricted procedure	8.2%	0.4%	24.6%	66.8%	5300
Negotiated procedure	28.4%	0.2%	11.8%	59.7%	40163
Accelerated negotiated procedure	39.8%	0.7%	29.5%	29.9%	3010
Competitive dialogue	17.1%	0.0%	9.9%	72.9%	2206
Negotiated without a call for competition	26.0%	0.1%	56.0%	17.9%	34648
Award of contract without prior publication of a contract notice	25.4%	0.1%	23.0%	51.6%	48672
Not specified	55.9%	4.9%	6.2%	33.0%	5490
Total	259541	3039	290208	863824	1416612

In table 6 we see that for the most frequently used procedure, the open procedure, around 20% of the CAs have only 1 bid.

In table 7 the education sector seems to have a special distribution with many contracts that have received just 1 bid. Construction and real estate is the industry with the smallest share of 1-bid contract awards.

Table 7: Distribution of bids by *selected* 2 digit CPV codes.

	missing	0	1	more than 1	total
Agriculture and Food	17.8%	0.2%	16.9%	65.2%	92203
Computer and related services	16.0%	0.2%	20.6%	63.2%	68934
Construction and real estate	22.2%	0.2%	7.4%	70.2%	254167
Education	12.0%	0.4%	41.1%	46.5%	44336
Energy and related services	18.6%	0.3%	20.8%	60.3%	27289
Environment and sanitation	17.3%	0.2%	13.2%	69.4%	62956
Financial and related services	18.9%	0.3%	19.9%	60.9%	30399
Health and social work services	19.5%	0.2%	21.1%	59.3%	28201
Other	17.1%	0.2%	25.5%	57.2%	711728
Research and development	13.5%	0.2%	15.5%	70.7%	5716
Transport and related services	22.5%	0.4%	16.6%	60.4%	90683
Total	259541	3039	290208	863824	1416612

Note: The selection of 2 digit CPV code is the one found when clicking on the first web page of the online version of TED.

4. Digging deeper.

From the overview tables in section 3 we now turn our attention to the CAs with just 1 bid in an attempt to characterize such contract awards even further.

4.1 One-bid all countries.

In this section we will focus on just two characteristics of the 1-bid CAs. First, in table 8 we look at the distribution of 2 digit CPVs across countries. In this table we show 2 digit CPV where at least one of the selected countries (the same selection as in table 2) has a percentage of at least 5. All the selected

countries have percentages above 5 in 2 digit CPV 33 (Medical equipments, pharmaceuticals and personal care products) but even within the CPV the percentages for each country vary quite a lot. The largest percentage is observed for Poland with Romania closely after. Countries like the UK , Denmark and France have much lower percentages. For the rest of the CPVs often only one of two of the countries show values of above 5%. Some of these 'single case', though show quite large percentages (CVP 34 for Germany, CPV 60 for the UK, CPV 66 for IT and the UK, CPV 80 for Poland, CPV 85 for the UK and CPV 90 for Denmark).

Table 8: Distribution of 2 digit CPV industry by country. Percentage points by column.

CPV	BG	DE	DK	ES	FR	IT	PL	RO	UK	other
33	44.51	10.53	24.72	19.68	12.55	40.64	62.84	52.48	8.36	25.61
34	4.28	13.98	5.53	4.42	6.32	4.01	1.00	2.58	2.42	3.82
38	1.88	7.36	6.21	4.22	1.68	3.23	3.26	2.11	4.46	5.78
50	6.50	2.21	1.58	7.03	5.17	3.74	1.76	5.64	4.30	4.25
60	2.03	5.03	1.81	7.35	5.79	2.52	0.45	0.87	17.09	2.10
66	1.37	0.53	5.98	1.87	3.77	12.27	0.99	1.34	9.99	2.10
79	4.50	1.84	3.84	4.96	3.63	1.84	1.23	2.60	6.06	3.94
80	0.79	1.43	1.58	0.49	4.27	1.14	9.47	0.25	2.81	2.58
85	1.91	1.39	1.02	1.73	1.51	4.50	1.78	2.32	10.61	2.15
90	2.63	3.99	12.53	3.85	5.27	5.73	1.81	2.42	2.11	3.05
Other	29.61	51.72	35.21	44.40	50.04	20.40	15.43	27.40	31.79	44.62
Total	7497	10387	886	5662	33645	11004	159339	11634	4306	45848

Notes: 33- Medical equipments, pharmaceuticals and personal care products; 34- Transport equipment and auxiliary products to transportation; 38- Laboratory, optical and precision equipments (excl. glasses); 50- Repair and maintenance services; 60- Transport services (excl. Waste transport); 66- Financial and insurance services; 79- Business services: law, marketing, consulting, recruitment, printing and security; 80- Education and training services; 85- Health and social work services; 90- Sewage, refuse, cleaning and environmental services.

In table 9 we study the distribution of one-bid contracts across selected countries (same as in table 2) and authority type. In table 9 we notice that authority types 'regional or local authority' (3), 'body governed by public law' (6) or 'other' (8) for many countries hold large percentages. 'Utilities' (4) and

'European Institution/Agency or International Organisation' (5) hold quite low percentages as do 'National or federal Agency/Office' (N) and 'Regional or local Agency/Office'(R).

.Table 9: Distribution of one-bid contracts across selected countries and authority type. Percentage points by row.

	1	3	4	5	6	8	N	R	Z
BG	17.30	11.00	6.16	0.01	60.68	3.40	1.37	0.05	0.01
DE	10.61	34.05	2.49	0.47	25.86	22.43	2.38	0.62	1.09
DK	12.87	61.63	3.16	1.02	13.66	4.18	0.11	2.71	0.68
ES	7.29	41.33	9.89	1.20	14.96	18.30	0.19	5.25	1.59
FR	6.05	30.42	1.29	0.10	19.87	13.18	0.33	0.59	28.17
IT	4.71	38.35	7.70	1.16	25.93	16.34	0.21	4.15	1.45
PL	4.00	7.30	1.67	0.01	49.57	35.62	0.37	0.70	0.76
RO	32.26	10.28	7.60	0.09	13.15	35.53	0.70	0.38	0.01
UK	11.15	47.12	1.30	0.46	30.26	5.43	0.88	2.44	0.95
other	19.41	15.76	4.43	1.93	36.45	13.47	3.58	3.20	1.78
Total	24988	43790	8213	1220	116277	77187	2842	3777	11914

Note: Codes in the Column headings are TED codes. Cells with a grey shade have percentages larger than 15. No shading in the 'other' row.

There are quite a few other variables that may deserve attention in this section but we have chosen to limit out preliminary study to the characteristics covered by the variables of table 8 and 9.

4.2 One-bid, Poland.

Finally, we will focus briefly on the one-bid contracts of Poland because of the very large number (and share) of contracts for this country.

In table 10 we show the distribution of selected 2 digit CPV codes against the procedure types. We observe a very large share of 1-bid cases for the combinations: 'Medical equipments, pharmaceuticals and personal care products' (33) and 'Laboratory, optical and precision equipments (excl. glasses)' (38) versus open procedure. Another combination that stands out is 'Business services: law, marketing, consulting, recruitment, printing and

security' (79) with a quite low share of open procedure cases and a somewhat larger share of restricted procedures compared to the other CPVs.

Table 10: Distribution of selected 2 digit CVP codes against procedure type (TED definitions) for Poland.

	1	2	3	4	6	C	T	V	Z
33	99.38	0.12	0.16	0.01	0.02	0.00	0.27	0.03	0.00
34	88.39	5.08	3.83	0.19	0.00	0.00	1.88	0.63	0.00
38	98.44	0.13	0.21	0.10	0.00	0.00	0.83	0.29	0.00
50	87.29	1.67	1.21	0.78	0.07	0.04	4.67	4.27	0.00
60	90.25	1.67	0.42	1.25	0.00	0.00	3.48	2.92	0.00
66	92.23	2.80	0.83	1.15	0.06	0.13	2.36	0.45	0.00
79	58.02	22.92	0.97	0.36	0.31	0.05	10.34	7.03	0.00
80	82.32	0.09	1.84	0.04	0.91	0.00	0.94	13.85	0.00
85	87.50	0.18	0.14	8.02	0.00	0.00	3.00	1.17	0.00
90	87.08	1.11	0.10	0.38	0.56	0.00	5.56	5.21	0.00
Other	81.23	2.64	0.37	0.74	1.12	0.08	7.81	5.98	0.02
Total	149074	1464	671	502	462	23	3051	4087	5

Notes: 33- Medical equipments, pharmaceuticals and personal care products; 34- Transport equipment and auxiliary products to transportation; 38- Laboratory, optical and precision equipments (excl. glasses); 50- Repair and maintenance services; 60- Transport services (excl. Waste transport); 66- Financial and insurance services; 79- Business services: law, marketing, consulting, recruitment, printing and security; 80- Education and training services; 85- Health and social work services; 90- Sewage, refuse, cleaning and environmental services. For the TED codes for procedure, please see first column in table 11 below.

Table 11: Distrubution of procedure type by contract type.

Procedure type	SERVICES	SUPPLIES	WORKS
1 Open procedure	21.56	78.05	0.39
2 Restricted procedure	76.16	22.54	1.30
3 Accelerated restricted procedure	57.53	41.88	0.60
4 Negotiated procedure	85.46	11.75	2.79
6 Accelerated negotiated procedure	91.77	6.06	2.16
C Competitive dialogue	56.52	30.43	13.04
T Negotiated without a call for comp.	73.52	23.34	3.15
V Award without prior contract notice	84.29	12.06	3.65
Z Not specified	100.00	0.00	0.00
Total	40199	118261	879

In table 11the distribution of procedure type across contract type is found. The combination of open procedure versus Supplies is a very frequent one as are the combinations Negotiated procedure, Accelerated negotiated procedure and Award without a prior notice against Services.

5. Conclusion.

We have in this study chosen to focus on contracts that have a potential to be classified as 'lack'of'competition' contracts due to them having received only one bid. Our descriptive approach has demonstrated that there are large variations in frequencies of one-bid contracts when it comes to countries, type of authority, CPV industries and type of procedure. Clearly this is a field that deserves more attention in future research.

6. References

Commission (2011): Evaluation Report Impact and Effectiveness of EU Public Procurement Legislation - Part 1, Staff Working Paper, SEC(2011) 853 final, p. xi.

la Cour, L. and A. Milhøj (2013): Public procurement – an explorative study of the contracts of the Tender Electronic Daily (TED). In Symposium i Anvendt Statistik, pp. 178-194.

la Cour, L., Milhøj, A. and Ølykke, G.S. (2015): Transparency Notices in the EU Public Procurement Regime: An Empirical Study of the use of Transparency Notices in Denmark, Sweden and the United Kingdom. *Public Procurement Law Review*, 5, pp 164-192.

http://ted.openspending.org/

A Giant leap for business statistics

Søren Kristensen, Danmarks Statistik

Skr@dst.dk

Indledning

Der har i de senere år været en tendens til at nationale statistikbureauer opretter særlige enheder til at tage sig af de største virksomheder i økonomien. Selvom disse enheders opgaver kan variere meget i både omfang og ambition, så er der den fællesnævner, at man forsøger at behandle data på tværs af de vigtigste statistikområder for en lille del af populationen. Det er på mange måder et brud med den tilgang, der traditionelt har været til arbejdet i de nationale statistikbureauer, hvor hvert statistikområde har haft ansvaret for sin egen population.

Man kan sige, at hvert statistikområde har eksisteret som sit eget økosystem, - eller "silo" som er den foretrukne metafor – og har kunnet udvikle sig mere eller mindre uforstyrret. Det har selvfølgelig nogle fordele, bl.a. at de personer, der arbejder med statistikkerne, har detailviden om det de skal måle og samtidig har mulighed for direkte kontakt til de store virksomheder. Ulemperne ved den statistiske bio-diversitet er selvfølgelig, at man risikerer manglende konsistens mellem statistikker, der burde korrelere eller på anden måde hænge sammen, at virksomhederne kan blive kontaktet mange flere gange end måske er nødvendigt, og at man også mister det samlede overblik over en virksomheds aktiviteter fordi indsamlingen er koncentreret om enkeltaspekter.

Sikringen af konsistens i data er dog langtfra den eneste grund til at forsøge at koncentrere arbejdet om de største virksomheder i én enhed. Der er også en voksende erkendelse af, at vi som statistikere er nødt til at forstå de store virksomheder og deres forretningsmodeller langt bedre end vi gør i dag. Uden denne forståelse kan det være svært for os at lave retvisende statistikker. En helt basal men afgørende problemstilling er her, hvad vi forstår som en virksomhed, eller rettere hvad vi forstår som vores enhed i erhvervsstatistikken.

Hvad er en virksomhed?

Vi har måske alle på et eller andet plan en ide om, hvad en virksomhed er; I en produktionsvirksomhed er der nogle bygninger, en lagerhal, en administrationsbygning, folk der kører med gaffeltruck, nogle der står ved nogle maskiner og andre, der går rundt med et clipboard for at kontrollere. Og på en eller anden måde er de alle del af den samme virksomhed.

Men det vi kan se, når vi arbejder med statistik er ikke nødvendigvis det, der svarer til virksomheder. Vores udgangspunkt er administrative registre - såsom CVR. Alle der kender noget til erhvervsstatistik ved, at virksomheder, og især store virksomheder, kan bestå af flere CVR numre, og at de kan være oprettet til forskellige formål. Der kan være oprettet en virksomhed, der ejer lagerbygningen. Administrationen kan også have sit eget CVR nummer, og de ansatte i virksomheden er måske tilknyttet et helt tredje CVR nummer. Hvis man bare vælger at se på de juridiske enheder, vil man så se et administrationsselskab, et ejendomsselskab og et vikarbureau og derved få et fragmenteret billede af denne tænkte virksomhed og dens aktiviteter.

Det er da heller ikke meningen, at man i erhvervsstatistikken skal bruge den juridiske enhed (altså det der svarer til CVR nummeret), men derimod skal man bruge det vi kalder den økonomiske enhed (på engelsk enterprise), som er en statistisk enhed, der er afledt af den juridiske enhed. Den økonomiske enhed er netop konstrueret for at man kan sikre et retvisende billede af en virksomhed og man ikke er bundet af den måde en virksomhed har valgt at organisere sig på rent juridisk. I en række erhvervsstatistikker har det imidlertid i praksis vist sig, at man har brugt den juridiske enhed i stedet for en økonomisk enhed. Det hænger primært sammen med at oversættelsen fra de juridiske enheder i det administrative system til statistiske enheder til brug for statistikproduktion er meget vanskelig og tidskrævende.

I Danmarks Statistik er vi netop begyndt på arbejdet med at få bedre viden om de største virksomheder i dansk økonomi. Vores sigte er både at få bedre konsistens i data og at sikre at vi får en bedre forståelse for, hvordan virksomhederne er organiseret så vi kan "danne" økonomiske enheder og dermed forbedre kvaliteten af statistikkerne.

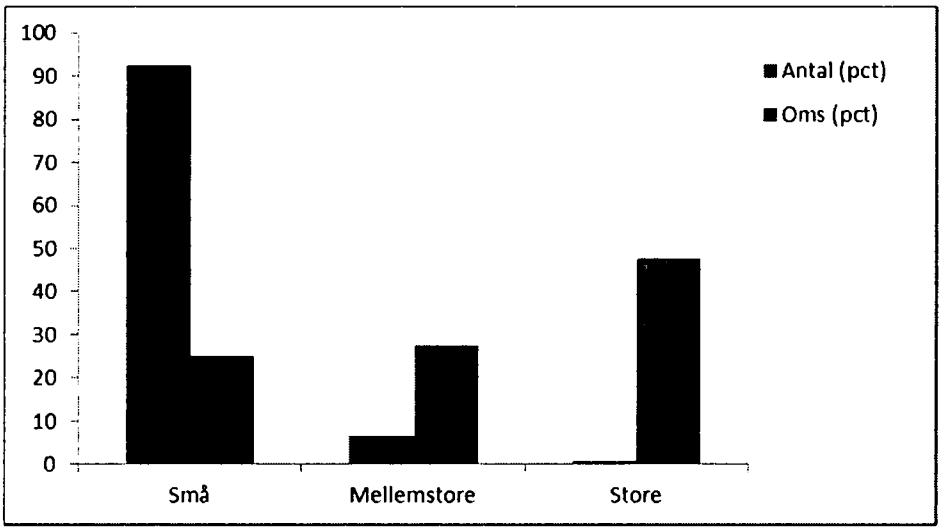
I dette paper vil jeg ikke komme så meget ind på konsistensaspektet, som i sig selv er meget vigtigt, men især fokusere på problemstillingen om, hvordan vi forstår en virksomhed og hvordan vi vil arbejde med økonomiske enheder. Der er ingen tvivl om, at den samlede gennemførelse af dette vil få ret markante virkninger for nogle statistikker, ligesom man også må forvente en ganske stor forskydning mellem brancher.

Aktiviteten koncentreret på få enheder

Et af de særlige kendetegn ved erhvervsstatistik er, at man stort set altid har at gøre med meget skæve fordelinger. Hvis man ser på den samlede økonomi, så er en meget stor del af omsætningen samlet på få virksomheder (her i praksis juridiske enheder), og det samme billede gør sig gældende, når man går ned i detaljen og ser på enkeltbrancher.

Som figuren neden for viser, så udgør de store virksomheder blot 1 pct. af den samlede population af virksomheder, men tegner sig for næsten halvdelen af den samlede omsætning. Statistikken her er reelt baseret på juridiske enheder, så hvis man så på økonomiske enheder ville koncentrationen værre større.

Antal firmaer (pct.) og omsætning (pct.) fordelt efter virksomhedsstørrelse 2013.



Kilde: Den generelle firmastatistik. Små virksomheder (0-9 ansatte), mellemstore virksomheder (10-99 ansatte), store virksomheder (100+)

Der er ikke som sådan noget nyt i, at man fokuserer på de store virksomheder i statistikproduktionen. Alle statistikområder har en liste af virksomheder, som de skal have svar fra for at undgå, at deres statistik bliver for usikker. Selvom man selvfølgelig kan imputere eller opregne, så er det alligevel de færreste, der er trygge ved det, når der er tale om meget store virksomheder, hvor selv små udsving kan forplante sig til offentliggjorte tal. Man har også i nogle tilfælde selv håndteret problemstillingen vedr. enheder, når man er blevet opmærksom på, at det var nødvendigt med data fra flere CVR numre for at kunne beskrive den af virksomhedens aktiviteter, man indsamlede data for.

Det der er afgørende nyt ved at organisere en enhed, der skal tage sig af de store virksomheder, er at man går fra en situation, der har været statistikspecifik og lidt ad hoc præget, til en situation hvor arbejdet bliver generelt og systematisk. Med fare for at bruge et lidt forslidt ord kan man også sige, at arbejdet vil blive mere proaktivt end før, hvor man typisk har taget sig af den type problemstillinger, når man opdager store udsving i data.

Hvorfor er arbejdet med store virksomheder kommet i fokus?

Som kort nævnt i indledning er Danmark Statistik langt fra den eneste statistikinstitution, der har valgt at gå den vej – og vi er bestemt heller ikke de første. Men der er ingen tvivl om, at det bliver mere og mere populært at organisere sig på den måde. Men hvorfor er det, at arbejdet med de store virksomheder har taget fart de senere år?

Et af de forhold der ofte nævnes, er at globaliseringen har medført, at det i stigende grad er svært følge og registrere virksomhedernes aktiviteter. Globaliseringen er jo langt fra et nyt fænomen og man tør jo næsten ikke bruge ordet mere fordi det er blevet lidt slidt. Men der er alligevel sket en del bare i den seneste ti års periode med åbningen af BRIK markederne og ikke mindst med den teknologiske udvikling som medfører, at kommunikation er meget nemmere billigere end før. Inden for det europæiske område har åbningen af Østeuropa, særligt i forbindelse med EU udvidelsen, medført, at der er blevet et helt nyt lavtløns arbejdsmarked til rådighed for virksomheder i Vesteuropa. Dette har uden tvivl haft en effekt på den globale værdikæde, som er blevet mere uigennemskuelig og ikke transparent for udenforstående.

I den traditionelle version bestod værdikæden i, at der blev oparbejdet råvarer i u-landene, som så blev transporteret til I-landene hvor de blev forarbejdet for derefter at blive solgt både i I- og U-lande. Senere kom udflytning af produktionsarbejdspladser til lavtlønsområder, som er en proces, der stadig er i gang. I dag er også administrationsopgaver, forskning og udvikling, og servicevirksomheder også noget der kan flyttes rundt. De store virksomheder – eller koncerner som det jo retteligt er – er i stigende grad blevet uafhængige af den nationale inddeling af kloden. Tværtimod kan de bruge de lovgivningsmæssige forskelle til at optimere deres situation ved at "flytte brikkerne rundt" på den rigtige måde.

En anden meget væsentlig grund til at arbejdet med de store virksomheder igen er kommet på dagsorden i statistikbureauerne er, at det europæiske statistikbureau – Eurostat – har presset en del på og medfinansieret arbejdet med at danne et overblik over koncerner i Europa. Tanken er, at man vil etablere et fælleseuropæisk koncernregister, som allerede nu eksisterer i en prototype, og at dette skal danne grundlag for en bedre statistik på Europæisk plan. Det man ønsker, er at kunne lave europæisk statistik, som ikke blot er en sum af de nationale statistikker. Argumentet er, at man får et forkert indtryk af den europæiske aktivitet ved blot at slå tallene sammen fordi en del af den aktivitet der er i de store koncerner er intern handel, som egentlig bør konsolideres.

Det er lige præcis også det problem – og det argument – der kan bruges i forhold til den nationale statistik og de store koncerner, der agerer inden for det nationale område. Koncerner består pr. definition af mere end én juridisk enhed og i mange tilfælde er der intern handel mellem de juridiske enheder. Hvor stor denne handel er ved vi ikke før vi har information fra koncernerne selv. Og de fleste af de koncerner, som vi er interesserede i at belyse nærmere, er meget store og agerer globalt. Det vi ser i Danmark, og kan måle i de danske statistikker, er kun et udpluk af deres aktiviteter. Hvis der bare var tale om simple input – output aktiviteter i forhold til udlandet kunne man måske nøjes med blot at forstå den danske del af en koncern.

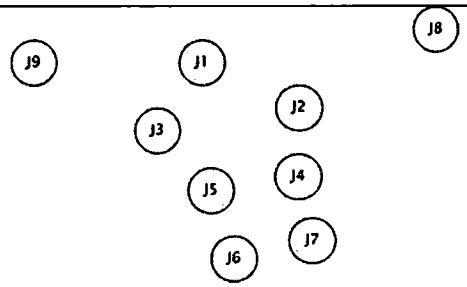
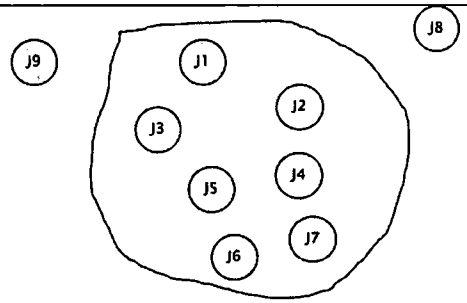
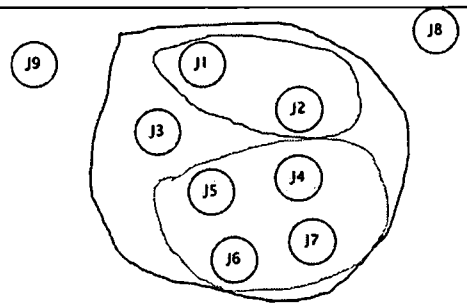
Men det er ikke helt så simpelt – der er strømme frem og tilbage, og der er aktiviteter der foregår i udlandet, men har direkte effekt på danske opgørelser. F.eks. betragtes produktion i udlandet som dansk produktion, hvis det er en dansk virksomhed, der ejer råvarerne, der indgår i produktion. En anden type aktivitet er merchanting, som beskriver den situation, hvor en virksomhed køber en vare i udlandet og sælger den videre til en anden virksomhed i udlandet. Her er der ikke aktivitet i Danmark, men køb og salg vil fremgå hos virksomheden og på betalingsbalancen.

Ved at danne en organisatorisk enhed, der skal tage sig af de store virksomheder, vender man statistikproduktionen lidt på hovedet. Hvor man før "nøjedes" med at indsamle meget specifikke informationer til brug for specifikke statistikker, lægger den nye model nu op til, at man skal have den bredere forståelse for koncernen på plads for bedre at kunne tolke data. Det stiller naturligvis også krav til vores kompetencer som statistikere. Vi er nødt til at opbygge en endnu større viden om de erhverv vi vil beskrive, end vi tidligere har været vant til.

Profilering – arbejdet med at forstå koncernerne

Det konkrete arbejde med at forstå koncernerne omtales som profilering i den statistiske verden. Profilering består i virkeligheden af to elementære trin. Det ene er at kortlægge en koncern, dvs. at kortlægge hvilke juridiske enheder, der er en del af koncernen, hvad de laver, og hvordan de evt. er forbundne. Det andet trin er at finde ud hvordan man kan danne økonomiske enheder i vores erhvervsstatistiske register ud fra de juridiske enheder, der indgår i koncernen. I første omgang vil vi dog primært overveje økonomiske enheder inden for de danske grænser, men det er klart at man kan i princippet danne dem på tværs af landegrænser.

Profilering: en skematisk oversigt over trinene

Et udsnit på 9 juridiske enheder, hvor vi ikke har viden om deres relationer	
Trin 1: kortlægning af koncernen. Her har man fundet ud af, at J1-J7 alle hører til samme koncern.	
Trin 2: danne økonomiske enheder. Her har man identificeret tre økonomiske enheder, hvoraf to er komplekse.	

Hvad er en økonomisk enhed, og hvordan adskiller den sig fra den juridiske?

Den økonomiske enhed er først og fremmest en meget central statistisk enhed, som skal anvendes i en lang række erhvervsstatistikker. Det er bestemt i den europæiske enhedsforordning fra 1993, som omtaler alle relevante statistiske enheder, og det står i de forskellige statistikspecifikke forordninger. Vi skal som nævnt tidligere ikke lave statistik på juridiske enheder, men økonomiske enheder (eller andre statistiske enheder).

Lidt omskrevet står der i definitionen;

- at en økonomisk enhed består af den mindste kombination af juridiske enheder, som tilsammen udgør en organisatorisk enhed, der producerer varer og tjenesteydelser, og som i et vist omfang har fri beslutningsret med hensyn til anvendelsen af sine ressourcer i den daglige drift.

Definition er om end ikke uklar så dog lidt svær at operationalisere. Måske er det særligt det helt centrale princip om fri beslutningsret, der er vanskeligt at afgrænse. Man kan sige som udgangspunkt, at antallet af økonomiske enheder vil være mindre end antallet af juridiske enheder, fordi der altid vil indgå en eller flere juridiske enheder i en økonomisk enhed. Der kan også være situationer, hvor en juridisk enhed bør splittes op for at man korrekt kan allokere dens aktiviteter til forskellige økonomiske enheder inden for koncernen.

Med disse muligheder er der tre grundtyper af økonomiske enheder:

- en økonomisk enhed består af én juridisk enhed
- en økonomisk enhed består af flere juridiske enheder
- en økonomisk enhed består af en eller flere juridiske enheder samt en andel af en juridisk enhed, der er splittet op.

Langt de fleste økonomiske enheder vil være af den første type. De komplekse økonomiske enheder – type to og tre – er der naturligvis meget færre af.

Hvad er reglerne for dannelse af komplekse økonomiske enheder?

Der er formuleret en række operationelle regler for hvornår man kan danne komplekse økonomiske enheder. Altså i hvilke situationer det kan være relevant at danne en økonomisk enhed af et antal juridiske enheder, som har relationer til hinanden.

En typisk situation er den hvor en virksomhed har valgt at udskille produktionsfaktorer i særskilte CVR numre. Det er f.eks. ganske normalt, at man opretter et selskab, der står som ejer af virksomhedens ejendomme eller der oprettes et selskab der ansætter medarbejderne og udlejer dem til produktionsvirksomheden. Det kan der være forskellige grunde til, f.eks. risikospredning eller der kan være nogle ansættelsesretlige grunde til det.

Hvis disse selskaber alene har til formål at betjene produktionsvirksomheden med leje af lokaler og personale, vil man som udgangspunkt betragte dem alle tre som dele af en kompleks økonomisk enhed. Det vil i praksis betyde, at man vil eliminere den interne handel mellem de tre selskaber, hvilket vil have flere konsekvenser for statistikkerne.

- Antallet af enheder vil blive mindre
- Aktiviteten i nogle brancher vil blive mindre
- For en række variabler vil den samlede aktivitet blive mindre – det gælder især omsætning og køb.

En anden typisk situation er, at en virksomhed har oprettet selskaber til at varetage forskellige hjælpefunktioner i forhold til det, der er dens kernefunktion. Det kan f.eks. være en produktionsvirksomhed, der udskiller IT eller bogholderiet som et selvstændigt selskab, eller hvad der er meget almindeligt, at der oprettes et salgsselskab som alene har til formål at sælge produktionsvirksomhedens varer på markedet. Fælles for denne type af aktiviteter er, at de ikke betragtes som en del af kerneforretningen, men netop som aktiviteter, der er nødvendige eller vigtige for at få kerneforretningen til at køre.

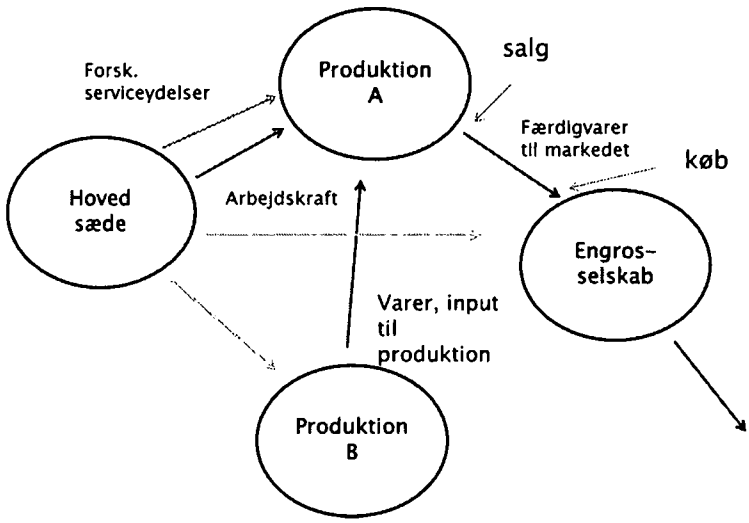
Som ovenfor gælder det, at hvis de enheder udelukkende servicerer produktionsvirksomheden så kan man betragte som den af en samlet kompleks økonomisk enhed. Konsekvenserne vil her også være, at der vil ske justeringer af forskellige statistikker og brancher. Hvis man f.eks. ville beregne en koncernomsætning ved at lægge omsætningerne fra alle enhederne sammen ville man få et for stort tal pga. den dobbeltregning der vil være.

En tredje situation er der hvor en virksomhed har valgt at oprette selskaber til de forskellige led i en værdikæde. Det omtales også som vertikal integration. Den situation kan opstå på flere måder, f.eks. ved at der oprettes et selskab til af produktionsprocesserne i en produktionsvirksomhed eller ved at man opkøber en underleverandør. Det samme gælder som ovenfor, at hvis selskaberne, der leverer til de forskellige stadier i processen alene producerer for den givne produktionsvirksomhed, så vil de som regel blive betragtet som en kompleks økonomisk enhed.

I praksis er eksemplerne ofte lidt mindre klare og lidt mindre skrivebordsagtige. Der er mange eksempler på hjælpeaktiviteter og vertikalt integrerede aktiviteter, der rent faktisk selv agerer på markedet og på den måde har lidt autonomi. Det betyder jo så, at det bliver sværere at vurdere, hvor grænsen skal gå for at inkludere en juridisk enhed i en kompleks økonomisk enhed, og at man under alle omstændigheder skal vurdere hvor stor en andel af omsætningen, der er intern.

Et tænkt eksempel

Neden for er et eksempel på hvordan en koncern og de interne strømme kan se ud. Det er et forholdsvist enkelt eksempel, der består af fire juridiske enheder; en produktionsenhed (A), som laver koncernens hovedprodukter, et hovedsæde som bl.a. udlejer arbejdskraft, en produktionsenhed (B), der laver mellemprodukter som indgår i hovedenhedens produktion, og endelig et engrosselskab, som sælger den producerede vare til markedet.



Enhed	Køb	Salg
Hovedsæde		100
Produktion A	300	1000
Produktion B		200
Engrosselskab	1000	1500
Total	**1300**	**2800**

Pga. de interne aktiviteter er der en del målbare strømme mellem de juridiske enheder. I tabellen er blot opgjort nogle få af dem. Hvis man ser på salgssiden, så er den reelle markedsværdi af produktionen 1500, men hvis man blot lægger salget fra de juridiske enheder sammen kommer man op på en sum på 2800. En stor del af denne forskel skyldes, at produktionen sælges internt inden den kommer ud på markedet.

Det er meget almindeligt, at produktionsvirksomheder er organiseret på denne måde, hvilket blot illustrerer, at der er en stor risiko for dobbeltregning, hvis man arbejder med juridiske enheder. Rent praktisk vil det ikke være en stor problemstilling i de data der indsamles, da statistikkerne er opmærksomme på at få de rigtige tal ind, men det kan være en problemstilling i f.eks. administrative data.

I det her konkrete tilfælde, vil det give mening at betragte de fire juridiske enheder som én økonomisk enhed jf. de regler, der er beskrevet ovenfor.

For yderligere at komplicere billedet, så er det jo også sådan at mange koncerner er opbygget på den måde, at de har centraliseret nogle typer af aktiviteter i nogle lande og at man herfra betjener resten af koncernen. Det kan f.eks. være bogholderiet, der ligger i Polen eller Skotland, eller produktionen, der ligger i Tjekkiet, mens forskning og udvikling ligger i Danmark. Det er klart, at det gør det kun vanskeligere at få et overblik over virksomhedens forretningsmodel, når de er organiseret på den måde. Og det gør det den del sværere at profilere dem fordi, der vil være mange enheder man set fra et nationalt perspektiv vil betragte som markedsorienterede, fordi den interne koncernhandel foregår med datter- eller moderselskaber i udlandet.

Hvad er målet – og konsekvenserne?

Behovet for at forstå koncernerne og evt. danne økonomiske enheder (og se data på tværs) er relevant for alle koncerner, men det er klart, at det reelt kun er meget få man vil kunne håndtere på en så intensiv måde i Danmarks Statistik. Vi har ikke fastlagt et konkret antal koncerner, der skal omfattes af dette arbejde, men det vil være de mest betydende koncerner udvalgt efter nogle nærmere kriterier, som vi også netop nu er i gang med at udfærdige.

Der er ingen tvivl om, at en øget fokus på økonomiske enheder som beskrevet ovenfor vil have en stor betydning for nogle brancher og statistikker. Det er for tidligt at give noget konkret bud på omfanget nu, men man kan i hvert fald komme med nogle generelle betragtninger om det.

Den helt overordnede konsekvens er, at det samlet set vil give et mere retvisende billede af den økonomiske aktivitet ved bl.a. at undgå dobbeltregning.

Men herudover kan man sige at:

- Antallet af enheder vil blive reduceret, pga. overgang fra juridiske til økonomiske enheder.
- At nogle brancher vil blive betydeligt mindre, herunder engrosbranchen, fordi de typisk vil blive defineret som hjælpeaktivitet.
- At den samlede aktivitet for nogle variablers vedkommende vil blive mindre. Dette skyldes at man ved dannelsen af økonomiske enheder konsoliderer variablerne.

Er SMV'er bare virksomheder, der ikke er blevet voksne?

Af

Mogens Dilling-Hansen

Department of Economics and Business Economics

Aarhus Universitet

Mail: dilling@econ.au.dk

1 Introduktion

Små og mellemstore virksomheder (SMV'er) dominerer dansk erhvervsliv, og derfor er det ikke overraskende at der er fokus på deres evne til at overleve, skabe jobs og skabe vækst. Diverse brancheforeninger er tilsyneladende alle enige om, at de største udfordringer de kommende ti år dels kommer fra fortsat outsourcing af produktion og dels kommer ved at erstatte jobs med maskiner. Væksten kan fortsætte, men ikke nødvendigvis med tilhørende stigning i beskæftigelsen, og der refereres til analyser, hvor det forudsiges at en stor andel af alle jobs vil være erstattet af maskiner de næste ti år[1].

Denne artikel vil ikke forsøge at rokke ved forventningerne til fremtidig erhvervsudvikling; men formålet er derimod at analysere SMV'er udvikling. Formålet er at identificere forhold, der er særlige for nystartede SMV'er og specielt undersøge om der er særlige hindringer for den omtalte automatiseringsproces.

Opbygningen af analysen er todelt. Den første del er fordelt over afsnit 2 til 3, hvor analyseapparat opstilles og anvendes til at undersøge om vækstmønstre er unikke for SMV'er. Denne del af analysen er baseret på register-data. Den anden del (afsnit 4) analyserer sammenhængen mellem virksomhedstype (SMV) og evnen til at omstille-virksomheden til at udnytte de positive effekter af automatisering, og denne del er baseret på mixed-method data fra et konkret automatiseringsforløb for SMV'er i fremstillingssektoren.

[1] Blandt andre refererer N. Milling m.fl., *Dansk Erhverv,* til en Oxford University undersøgelse, hvor 37 % af alle jobs erstattes af maskiner inden for 10 år, http://jyllands-posten.dk/aarhus/meninger/breve/ECE8192563-/Fremtidens-erhverv-er-kreative/

2 Teoretisk grundlag for SMV-analysen

Er SMV'er bare virksomheder, der enten ikke evner at vokse eller endnu ikke er blevet store virksomheder?

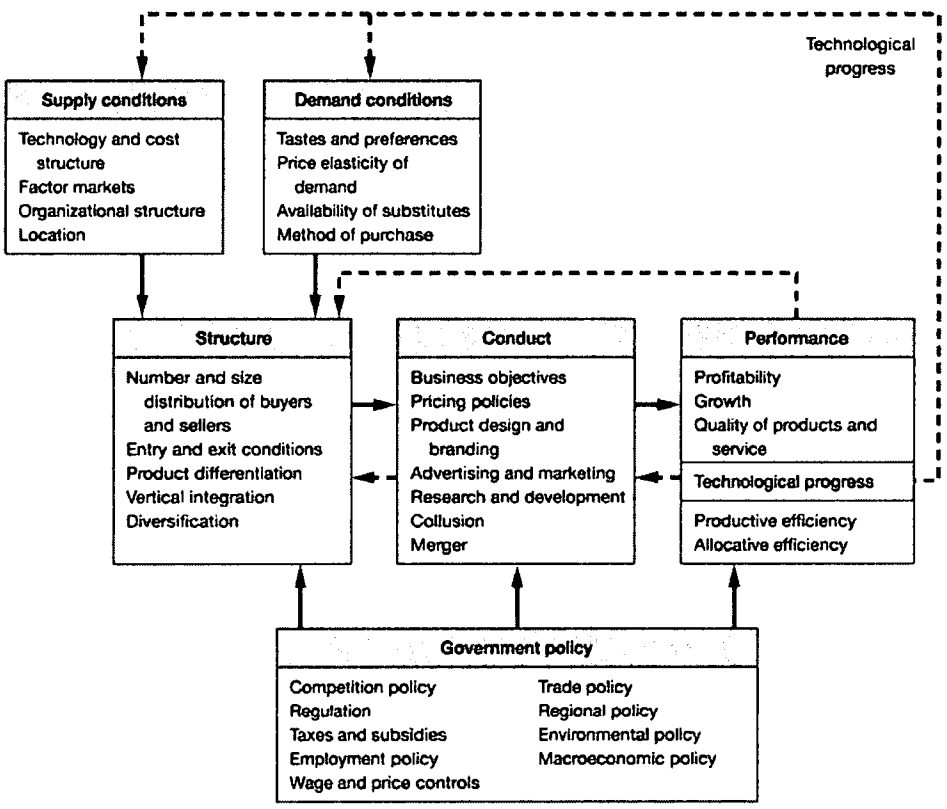
Der helt åbenlyse svar på det spørgsmål er, at spørgsmålet om en virksomheds størrelse afhænger af, hvorledes den klarer sig. Udtryk generelt er en virksomheds evne til at klare sig godt, dvs. performance, en funktion af en række faktorer, der både har med virksomhedens egne ressourcer (Conduct), med markedet (Structure) og med de generelle samfundsvilkår at gøre. Figur 1 viser en model baseret på Structure-Conduct-Performance paradigmet (S-C-P), og figuren understreger en række problemer ved at undersøge en virksomheds (f.eks. en SMV) performance:

- *performance kan ikke måles entydigt*
- *en virksomheds organisering har mange facetter*
- *markedsforhold påvirker en virksomhed på mange måder*
- *generelle markedsvilkår, offentlig regulering (love mv.) påvirker den enkelte virksomhed*

... og sammenhænge er ikke entydig (kausaliteten er ikke klar)

På trods af disse åbenlyse mangler er S-C-P paradigmet det fremherskende når virksomheder analyseres, dels på grund af manglende alternativer og dels fordi S-C-P tankegangen er velegnet til at lave partielle analyserer medfokus på virksomhedens performance uanset teoretisk tilgang.

Figur 1 Model baseret på Structure-Conduct-Paradigmet (SCP).



Kilde: Lipczynski m.fl.(2013)

S-C-P baserede modeller til forklaring er også et velegnet redskab til at forklare SMV'ers adfærd og evne til at tjene penge; men den konkrete hypotese om SMV'ers eventuelle generelle forskellighed fra større virksomheder er svær at teste.

Det skorter ikke på partielle analyser, der påviser forskelle i produktivitet, organisering og adfærd generelt; men litteraturen anfører generelt enighed om, at SMV'er ikke er "scaled-down firms" af to grunde.

For det første er der den fundamentale forskel i adgang til finansielle ressourcer, hvor store virksomheder af flere grunde er bedre til at skaffe adgang til ressourcerne – mindre kapitalanlæg, generel risikoprofil, sektorbaserede forskelle, forskelle i vækststrategier, generel vækstpotentiale, tilgang til profitmotiv og generelle skala-ulemper er fremført som væsentlige argumenter bag SMV'ers manglende evne (og lyst) til at blive større. Se Cressy & Olofsson (1997) for en oversigt over disse forskelle.

For det andet er SMV'ers evne og tilgang til knowledge management forskellige fra storevirksomheder. Igen er det ikke de manglende finansielle ressourcer, der som sådan skaber en anden måde at håndtere viden; men der er tale om fundamentale/strukturelle forskelle i håndtering af viden. Desouza & Awazu (2006) viser, at generel 'socialisering' anvendes til at håndtere "tacit knowledge" i større omfang, at generel viden deles i større omfang af alle medarbejdere, at "knowledge loss" håndteres fint på trods af nøglemedarbejderes viden, at eksterne kilder til viden i langt større omfang udnyttes og at mennesket i større grad end teknologien er i fokus ved håndtering af knowledge management. Viden håndteres efter forfatternes konklusioner på en mere human måde.

Et alternativ til S-C-P tilgangen til at undersøge SMV'er evne til at vokse er, alene at kigge på den faktiske vækst uden tilhørende forklaring af udviklingen. Denne tilgang er baseret på Gibrats lov, som grundlæggende blev formuleret som et statistisk fænomen, se Gibrat (1931). Den empirisk baserede model er senere døbt "*the Law of Proportional Effect*", fordi den antager uafhængighed mellem virksomhedens størrelse og vækstrate. Det betyder at (i) alle størrelsesopdelte klasser af virksomheder vil have samme proportionale vækstrate i gennemsnit og (ii) spredningen omkring fælles middelværdi er konstant over alle klasser. Denne relation er formaliseret af Sutton (1997) ved følgende beskrivelse af væksten fra periode *t-1* til *t*:

$$(1) \qquad X_t - X_{t-1} = \varepsilon_t X_{t-1}$$

som giver følgende sammenhæng mellem X_t og X_0

$$(2) \qquad X_t = (1+ \varepsilon_t) X_{t-1} = (1+ \varepsilon_t) (1+ \varepsilon_{t-1}) \dots (1+ \varepsilon_3) (1+ \varepsilon_2) (1+ \varepsilon_1) X_0$$

Med passende ikke-ekstreme værdier af tidshorisont og ε kan denne relation log-approksimeres ved

(3) $$\text{Log } X_t = \text{Log } X_0 + \varepsilon_t + \varepsilon_{t-1} + \ldots + \varepsilon_3 + \varepsilon_2 + \varepsilon_{1e)}$$

Med antagelsen om uafhængighed mellem ε_t og given middelværdi og fast varians er Log X_t log-normal fordelt for $t \rightarrow \infty$, og dermed kan modellen testes med simpel regressionsanalyse.

Geroski (1995) forsøger at tolke processen som en økonomisk læringsproces, hvor viden og erfaring (knowledge and experience) akkumuleres med samme hastighed; men denne tilgang har ikke vundet gehør. Derimod er modellen velegnet til at teste særlige grupper (klasser) for afvigelser fra Gibrats lov, hvor en signifikant $\beta_1 > 1$ medfører, at den udvalgte klasse har større vækst end gennemsnittet.

(4) $$\text{Log } X_t = \beta_0 + \beta_1 \text{ Log } X_0 + \varepsilon_t$$

I det efterfølgende afsnit er Gibrats lov testet på en særlig gruppe af SMV'er, nemlig iværksætter-virksomheder, som har særlig interesse for den fremtidige vækst, se f.eks. Rosa & Scott (1999).

3 SMV vækst og overlevelse

Udover den almindelige interesse for virksomheders evne til at overleve, så er der en væsentlig pointe ved at kigge på overlevelsesrater før Gibrats lov bliver testet. Anvendes model (4) til estimationerne, så er der to væsentlige krav til data

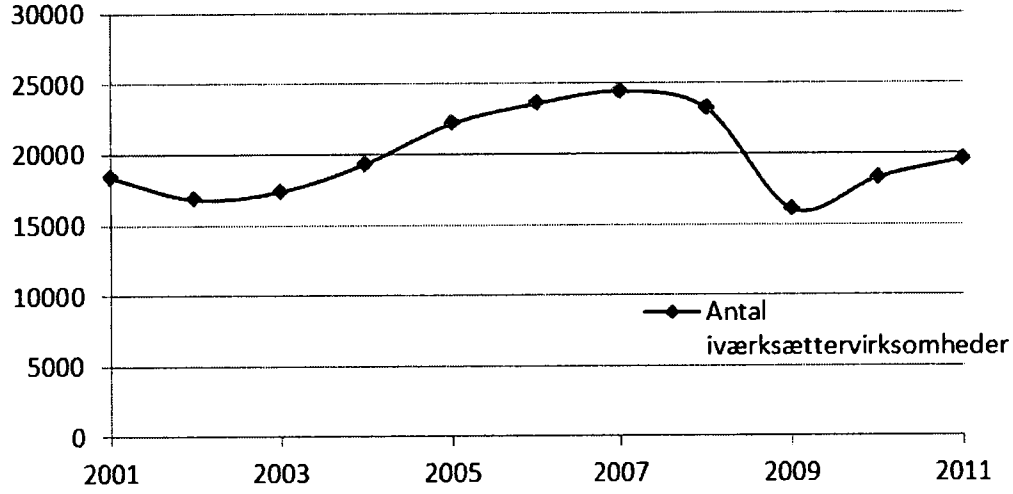
a) Performance skal måles på en valid måde
b) Virksomheden skal have tilgængelige data i hele analyseperioden

Performance burde måles ved omsætning, værditilvækst eller lignende, men for indeværende er kun antal beskæftigede til rådighed. Specielt for små virksomheder viser erfaringer generelt, at både opgørelse af antal beskæftigede og tolkningen er behæftet med nogen usikkerhed.

Analyserne i dette afsnit er baseret på udviklingen i iværksættere i Danmark for perioden 2001 til 2011, og figur 2 viser antallet af nye virksomheder i perioden. Omkring 20.000 nye virksomheder er årligt registreret som *iværksætter*virksomheder, medens Danmarks Statistik opgør det samlede antal *nye* virksomheder til ca.30.000 stk.

Der ses en klar pro-cyklisk udvikling i antallet af iværksættere. Finanskrisen gav et fald fra 2008 (23.182 iværksættere) til 2009 (16.047 iværksættere) på godt 30%, og det svarer nogenlunde til de estimerede fald i omsætningen for alle virksomheder.

Figur 2 Antal iværksættervirksomheder, 2001-2011

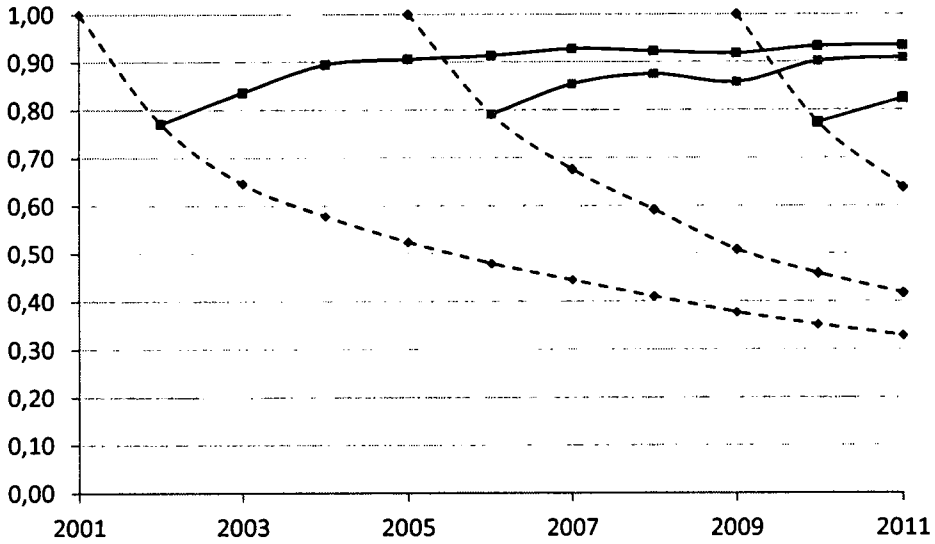


Kilde: Danmarks Statistik, register over iværksættere

Figur 3 viser to overlevelsessandsynligheder for tre kohorter af iværksættere startet i hhv. 2001, 2005 og 2009. På linie med Danmarks Statistiks overlevelsesrater (se f.eks. http://www.dst.dk/da/Statistik/NytHtml?cid=19637) er den kumulerede overlevelses-sandsynlighed monotont faldende over tid med den største risiko for exit det første år (ca. 30%) og ellerede efter tre års levetid falder sandsynligheden for exit til omkring 10%.

For kohorten af iværksættere, der startede i 2001 (samlet 18.391 stk), er der 6.057 virksomheder tilbage med et aktivt cvr-nummer (cvrnr). Der er naturligvis en del af disse virksomheder, der af forskellige grunde fortsætter med andet cvr-nummer (reorganisering, fusion mv.), men efter 10 år er andelen af overlevende virksomheder reduceret til omkring 1/3 af 2001-populationen. De estimerede modeller til test for proportional vækst (Gibrats lov) kan kun estimeres for (sub-)populationen af overlevende virksomheder (n=6.057).

Figur 3 Overlevelsessandsynligheder for nye virksomheder, 2001-2011



Noter: De stiplede linier er den kumulerede sandsynlighed for at overleve, vist for årgang 2001,2005 og 2009. Den fastoptrukne linie med firkantede symboler er betingede sandsynligheder for overlevelse vist for årgang 2001, 2005 og 2009. Kilde: Danmarks Statistik, register over iværksættere

De betingede sandsynligheder illustrerer denne proces ved at beskrive sandsynligheden for at overleve til periode *t* **GIVET** virksomheden allerede er *(t-1)* år gammel. I figur 3 er de betingede overlevelsessandsynligheder beregnet fra år 2 og det ses af figuren, at allerede efter 3 år er en iværksætters sandsynlighed for at opleve en exit på linie med den generelle risiko for exit – omkring 10%.

På nær en lidt større risiko for exit i året efter finanskrisen er både den kumulerede og den betingede sandsynlighed for overlevelse konstant over hele perioden, så selv om analyserne nedenfor kun er baseret på 1/3 af 2001-populationen, må det vurderes at det alvorligste problem med data formentlig er, at de første tre års frafald (omkring 42% af alle 2001-virksomheder) med stor sikkerhed er SMV'er der er forskellige fra større virksomheder, der har langt lavere risiko for exit.

Test af vækstmønster for SMV'er er i denne analyse foretaget på 2001 kohorten med en endelig population af 6.055 virksomheder, der har overlevet perioden 2001 til 2011. Den afhængige variabel i analyserne er 2011-beskæftigelsen[2] i virksomheden; omsætningstal eller tilsvarende ville være klart at foretrække, men valget er alene begrundet i tilgængelighed af variable. Resultaterne er præsenteret i (5) og (6) for hhv. alle virksomheder i 2001-kohorten, (5), og for større iværksættervirksomheder, (6). En større

[2] Den anvendte variabel er 'aarsvaerk', som er det samlede antal beskæftigede omregnet til fuldtidsansatte. Iværksætteren er i alle tilfælde lagt til aarsvaerk.

iværksættervirksomhed er i denne analyse defineret ved en fuldtidsbeskæftigelse på mindst 20 personer i start-året 2001.

Analyserne nedenfor er baseret på analyseenheden 'iværksættervirksomhed', hvilket vil sige at et unikt CVR-nummer har samme vægt uanset virksomhedens størrelse. Det betyder at fordelingen af virksomheder på størrelse er ekstremt højreskæv, hvilket kan se af den simple krydstabel i tabel 1. Tabellen opdeler de 6.055 iværksættervirksomheder efter beskæftigelsesstatus i 2001 og 2011: De 4022 virksomheder, der i 2001 havde højest én fuldtidsbeskæftiget person i virksomheden, svarer til ca. 2/3 af alle iværksættervirksomheder i 2001.

Tabel 1 Iværksættervirksomheder fordelt efter beskæftigelse, 2001-kohorte

	Beskæftigelse > 1 i 2011	*Højest én fuldtids-beskæftiget, 2011*	*Ialt*
Beskæftigelse > 1 i 2001	*1794*	*241*	*2035*
Højest én fuldtids-beskæftiget, 2001	*1449*	*2571*	*4022*
Ialt	*3243*	*2812*	*6055*

Estimationerne er baseret på en simpel lineær regression, og det ses af (5), at relationen er statistisk signifikant med en forklaringsgrad på ca. 35%. Tolkningen af estimatet på β_1 er interessant, fordi estimatet er under 1 og både signifikant forskellig fra 0 og 1. Gib rat's lov om proportionale vækstrater må ganske vist som forventet forkastes, men den forventede stigning i beskæftigelsen over den 10-årige analyseperiode er overraskende lav ... hvilket også kan ses af at den gennemsnitlige beskæftigelse kun stiger fra ca 2½ til 4½ fuldtidsbeskæftigede i perioden 2001 til 2011 for den gruppe af 2001-kohorten, der formår at overleve i hele perioden.

(5) $\text{Log } X_{2011} = 0{,}457 + 0{,}815 \text{ Log } X_{2001} + \varepsilon$ ***(2001-kohorte)***

Noter: n=6.055 observationer (2 iværksættervirksomheder er ikke SMV jf. definition: 50-250 ansatte). $Aarsvaerk_{2011}$ =4,31; $Aarsvaerk_{2001}$ =2,52. R^2-adj=0,35, $S.E(\beta_1)$ = 0,014, $P(F>3286,8) < 0,0001$.

(6) $\text{Log } X_{2011} = -0{,}11 + 0{,}908 \text{ Log } X_{2001} + \varepsilon$ **(2001, $besk_{2001}$>20)**

Noter: n= 34. $Aarsvaerk_{2011}$ = 27,6; $Aarsvaerk_{2001}$ =29,1. R^2-adj=0,04, $S.E(\beta_1)$=0,585, $P(F>2,41)$=0,131.

Den forventede positive udvikling i beskæftigelse er forventet i alle typer af S-C-P

baserede analyser, og det gælder specielt for de større virksomheder: Markedsvilkårene er kendte og virksomheden kender til dem (S), virksomhedernes interne ressourcer er givet begrænsede, men større end hos de helt små iværksættere (C), og derfor vil der forventes en klart højere koefficient til store virksomheder. Det er imidlertid ikke tilfældet for den – meget lille – population af større iværksættervirksomheder, der startede i 2001. I (6) er vist den insignifikante regressionsanalyse af de største virksomheder, hvilket i Gibrat-regí vil sige at der er tale om en random walk i beskæftigelsen, men ud fra alle økonomiske forventninger er dette resultat svært at tolke.

Resultaterne skal ses i lyset af at det er beskæftigelsen, der analyseres; men den insignifikante model præsenteret i (6) er både overraskende og interessant. Et bedre datagrundlag vil være fundament for videre analyser, der kan undersøge om det faktisk er et robust resultat, at størrelse af iværksættervirksomhed faktisk har en negativ effekt på fremtidig performance.

4 SMV'er og automatisering

Den anden del af denne analyse, der kigger på SMV'er evne til at vokse, har til formål at undersøge produktionsvirksomheder indstilling til at automatisere produktionsprocessen. Automatisering af produktionsprocessen er anset for at være en af nøglekomponenterne i SMV'er evne til at overleve, fordi automatisering betyder substitution af (dyr) arbejdskraft med kapital-intensive produktionsanlæg.

Argumentet bag automatisering er baseret på et forventet løft i produktivitet og/eller effektivitet ved automatisering; men specielt SMV'er kan have problemer med at gennemføre denne proces, fordi ressourcerne (både finansielle og knowledge-based) er begrænsede. Dette tema var i fokus for en større udviklingsprojekt (AUTO-projektet) gennemført af Dansk Produktions Univers (DPU), Hedensted, i 2014. 30 SMV'er fik tilbudt et individuelt automatiserings- og/eller digitaliseringsforløb, og det er denne proces, der er analyseret ved hjælp af et mixed-method approach, se Dilling-Hansen (2014)[3].

Motivet for deltagelse var for de 30 virksomheder, at de selv kunne bestemme hvilken type af automatiseringsforløb, der skulle iværksættes. Projekterne er efterfølgende inddelt i egentlige automatiseringsforløb og mindre ambitiøse automatiseringsforløb.

Resultaterne er baseret på et mixed method analyse-design, hvor der blev foretaget en kvantitativ spørgeskemabaseret undersøgelse FØR implementering af automatise-

[3] Papiret er en del af og finansieret af projektet "Automatisering og Digitalisering i små og mellemstore produktionsvirksomheders produktionsprocesser" (Konkurrenceudsatte midler, Den Europæiske Socialfond, ESFK-13-0048).

ringsprojektet og et kvalitativt dybde-interview EFTER implementeringen af projektet. Analyserne og projektet blev afsluttet med rapporten i december 2014.

Der var to meget klare motiver for virksomhedernes deltagelse. Tabel 2 summerer dette resultat ved at karakterisere de mere driftsorienterede automatiseringsprojekter som digitaliseringsprojekter (ca 1/3), medens egentlige automatiseringsprojekter var dem, der både skulle øge produktivitet og effektivitet (ca 2/3). Dimensionen 'innovativ' er bestemt ud fra analyser af virksomhedernes nuværende og planlagte teknologiske stade.

Tabel 2 Innovation og grad af automatisering

Virksomhedtype	Grad af digitalisering	Grad af automatisering
Ikke innovative	72,5%	37,5%
Innovative	39,6%	70,4%

Kilde: Dilling-Hansen (2014)

Resultaterne i Dilling-Hansen (2014) er analyseret ud fra den viden, at der må forventes en klar self-selection bias i virksomhedernes tilgang og vurdering af automatiseringsprojektet (derfor valget af mixed method tilgangen), og effektvurderingen er meget positiv, selvom det primære formål (at implementer state-of-the-art teknologi) virker noget ambitiøst med den relativt korte tidshorisont.

De fundne resultater er meget i tråd med Gunasekaran m.fl. (2001), der identificerer fire grundlæggende hindringer for implementering af ny teknologi:

i) Ressourcer (først og fremmest finansielle)
ii) Ledelsens ambitionsniveau (strategiplaner mv)
iii) Organisatoriske evner (ekspertise, samarbejde mv)
iv) Uddannelse og træning

Resultaterne er præsenteret i tabel 3, hvor den væsentligste drivkraft bag automatisering er ønsket om øget effektivitet. Det interessante ved dette resultat er, at der er et lige så klar sekundært motiv for at deltage i udviklingsforløbet, nemlig at opnå erfaringer med at samarbejde med eksterne partnere, medens 'best-practice argumentet' i langt mindre grad er en drivende kraft.

Tabel 3 Motiver for deltagelse i AUTO-projektet

	Alle virksomheder	Små virksomheder	Større virksomheder
Motiv for deltagelse i AUTO			
Vækst i omsætningen	30%	29%	31%
Best-practice teknologi	40%	41%	38%
Interne ressourcer/effektivisering	97%	94%	100%
Erfaringer med eksterne partnere	80%	71%	92%

Kilde: Dilling-Hansen (2014)

Spørger man virksomhederne direkte om de langsigtede mål, så er der en klar interesse i at vokse; men selv om SMV-status dermed kunne opfattes som en midlertidig status, så er fokus for udvikling og automatisering, at der findes et klart alternativ til at vokse i størrelse, nemlig at blive bedre til at udvikle virksomheden sammen med andre virksomheder i værdi-kæden.

To interessante detaljer ved de fundne resultater er, (1) at der blev fundet et klart fagligt udbytte af automatiseringsprojektet; men en del af denne forbedring faktisk var en justering af det opfattede teknologiske niveau for virksomheden, (2) at selv om der blev fundet klare forbedringer i proces- og performancestyring, så var denne proces langt fra afsluttet ved deltagelse i AUTO-projektet. Innovation tager tid!

5 Afrunding

Formålet med denne analyse er at undersøge SMV'ers evne til at vokse, ikke gennem fusioner men ved organisk vækst. Et særligt problem ved debatten om SMV'er tyngde i den danske erhvervsstruktur er, at det forventes at en stor del af dynamikken skabes gennem iværksætteri. Iværksættervirksomhedernes evne til at vokse er dog langt fra en simpel proces, og specielt bør de mange små nye iværksættere have større fokus i den danske erhvervspolitik.

Fornyelse gennem automatisering er derimod en meget frugtbar vej for SMV'er, så længe det huskes at automatisering ikke kun er en teknologisk opgave, men et projekt hvor både teknologisk viden, evner, vilje, uddannelse og en organisation gearet til samarbejde er vigtige faktorer bag succesfuld gennemførelse af automatiseringen.

Referencer

Cressy, R & C Olafsson (1997), "European SME Financing: An overview", *Small Business Economics* **9**: 87–96, 1997.

Desouza, K. C. & Y. Awazu (2006), "Knowledge Management at SMEs: Five Peculiarities", *Journal of Knowledge Management*, Vol. 10, No.1, 2006, pp. 32-43.

Dilling-Hansen, M. (2014), *Automatisering og digitalisering i små og mellemstore virksomheder*, Aarhus Universitet.

Geroski, P. (1995), "What do we know about entry?", *International Journal of Industrial Organisation*, 13, pp. 421-440.

Gibrat, R. (1931), *Les Inégalitiés Èconomiques*, Faculte de Droit , Université de Lyon, 1931.

Gunasekeran, A. H.B. Marri, R. McGaughey & R.J. Grieve (2001), "Implications of organization and human behaviour on the implementation of CIM im SMEs: An empirical analysis", International Journal of Computer Integrated Manufacturing, vol 14, pp. 175-185, 2001.

Lipczynski, J., J.Wilson & J. Goddard (2013), *Industrial Organisation*, 4th ed., Pearson.

Rosa, P. & M.Scott (1999), "The prevalence of multiple owners and directors in the SME sector: implications for our understanding of start-up and growth", *Entrepreneurship & Regional Development,* 11 (1999), pp.21-37.

Sutton, J. (1997), "Gibrat's Legacy", *Journal of Economic Literature*, vol. 35, no. 1, pp. 40-59

Towards A Theory of Social Data: Predictive Analytics in the Era of Big Social Data

Niels Buus Lassen, Ravi Vatrapu, Lisbeth la Cour, René Madsen, Abid Hussain, Copenhagen Business School

INTRODUCTION

In this chapter, we will advance a theory of social data that distinguishes between constituent dimensions of social graph (i.e., socio-technical affordances of social media networks) and those of social text (i.e., communicative and linguistic properties of social media interactions) as distinct but complementary elements of predictive big social data analytics. Additionally, to illustrate the validity and applicability of our proposed theory, we adhered to the schematic steps advocated by Shmueli and Koppius (2011) in building empirical predictive models that blend social graph analysis with social text analysis to: (1) compute correlations between social data from multiple social media platforms (i.e., Facebook and Twitter) and the financial performance (i.e., quarterly revenues) of corporate entities (i.e., iPhones and H&M), as well as; (2) make predictions about the future performance of these corporate entities. In doing so, we endeavor to provide an answer to the following research question: *How can big social data analytics be utilized to predict business performance?*

This paper comprises four sections, inclusive of this introduction.. In Section 2, we construct our theory of social data by extending Vatrapu's (2008, 2010) concepts of socio-technical affordances and technological intersubjectivity to the domain of social media. Section 3 outlines our methodological strategy for extracting and analyzing big social data to build empirical predictive models of business performance. Results from analyzing these empirical predictive models are also reported in Section 3. The last section, Section 4, summarizes the: (1) implications of this study to both theory and practice; (2) insights to be gleaned towards informing the application of predictive analytics to big social data; (3) possible limitations in the interpretation of our empirical findings, and; (4) probable avenues for future research.

TOWARDS A THEORY OF SOCIAL DATA

To bridge the knowledge gaps in extent literature, we advance a theory of social data that extends Vatrapu's (2008, 2010) concepts of socio-technical affordances and technological intersubjectivity to the domain of social media. Social media (e.g., Facebook and Twitter), at the highest level of abstraction, involve social entities interacting with: (a) technologies (e.g., an individual using the Facebook app on his/her smartphone), and; (b) other social entities (e.g., the same individual liking a picture of a friend on the Facebook app). Vatrapu (2008, 2010) labelled

these interactions as *sociotechnical interactions* (see also Vatrapu and Suthers 2010). Socio-technical interactions yield electronic trace data that we termed as ***social data***. To derive a theory for social data, we must first determine the constituents of socio-technical interactions. As acknowledged by Vatrapu (2010), socio-technical interactions are realized through: (a) a social entity's perception and appropriation of *socio-technical affordances*, as well as; (b) the structures and functions of *technological intersubjectivity* (Vatrapu 2010).

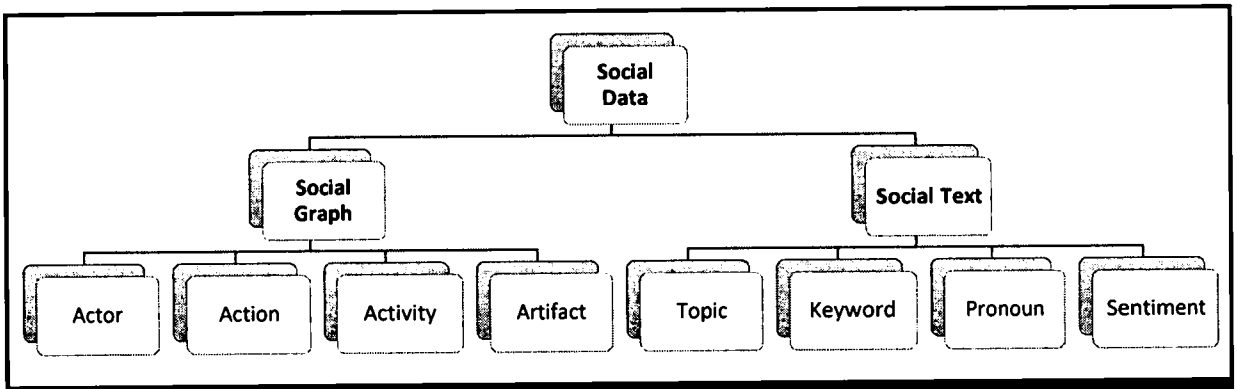


Figure 1: Theory of Social Data

As an illustration of our theory, consider the earlier example of an individual liking a friend's picture on the Facebook app. The performance of such a simple sociotechnical interaction already activates multiple social data elements: an *actor* (i.e., individual) performing an *action* (i.e., liking) on an *artifact* (i.e., Facebook app) for the purpose of expressing a *sentiment* (i.e., like) and contributing to a collective *activity* (i.e., expanding the social network timeline). Such micro social-technical interactions, when amassed in large volumes, constitute the macro world of big social data, the core premise of this paper.

METHODOLOGY AND ANALYTICAL FINDINGS

In this section, we presents details about the collection, preparation, exploration, selection, modelling and reporting of two big social data sets to illustrate different aspects of our proposed theory of social data. In general, we adhered to the methodological schematic recommended by Shmueli and Koppius (2011, p. 563) for building empirical predictive models. The remainder of this section is organized in accordance with Shmueli and Koppius's (2011) eight methodological steps of predictive model building as depicted in Figure 2.

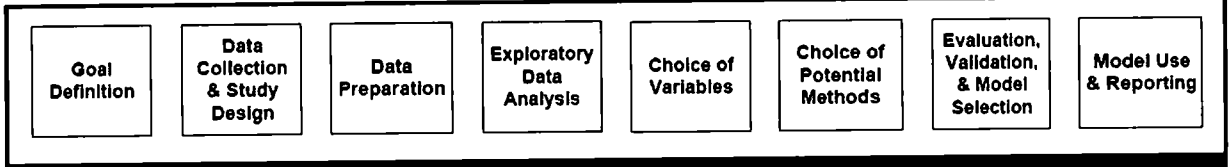


Figure 2: Methodological Steps in Predictive Model Building [Shmueli and Koppius 2011, p. 563]

Step 1: Goal Definition

Our primary goal was to build empirical predictive models of sales from big social data. More specifically, by applying predictive analytics to big social data, we strive to model and accurately predict the real-world numerical outcomes of quarterly sales of Apple iPhone & H&M revenues.

Step 2: Data Collection and Study Design

We discuss the rationale for the study design first followed by details on data collection.

Study Design: The study was designed to collect and analyze big social data sets that serve as illustrative case studies for predictive analytics. Therefore, we deliberately introduce variance into both the predicted variable of sales as well as the predictor variables of social data attributes.

With regard to the predicted variable of sales, we sought to incorporate variance in terms of *product types* (i.e., Apple iPhone: consumer electronics and H&M: fashion-clothes) and *sales channels* (i.e., offline and online; direct and retail). As for the predictor variables of big social data, we incorporated variance in terms of *social media platforms* (i.e., Facebook and Twitter), *theory of social data attributes* (Social Graph: actors, actions, artifacts and Social Text: keywords and sentiment), *dataset sizes* (few millions to hundreds of millions of data points), and *data time periods* (few months to years). Table 1 summarizes the characteristics of the two big social datasets that have been collected, processed and analyzed in this paper.

Table 1: Big Social Datasets Collected for Predictive Analytics

Company	Data Source	Time Period	Size of Dataset	Mapping to Social Data Attributes
Apple[1]	Twitter	2007 → June, 2015	500 million+ tweets containing "iPhone"	▪ **Social Text**: Keyword ("iPhone") ▪ **Social Text**: Sentiment
H & M[2]	Facebook	January 01, 2009 → March, 2015	~15 million Facebook events	▪ **Social Graph**: Actions (Total Likes) ▪ **Social Graph**: Artifacts (Posts and Comments) ▪ **Social Graph**: Actors (H&M + Non-H&M) ▪ **Social Text**: Sentiment

Data Collection: We now present details on the methods and tools used for data collection for the two big social datasets.

Twitter (Apple: "iPhone")

We collected over 500 million tweets containing the phrase "iPhone" in the period 2007-2015 (till March, 2015) via Topsy Pro Analytics[3]. Technically, our data collection did not connect to the Twitter firehose, but rely on a Twitter API solution with full access to all Twitter data.

[1] URL: https://www.apple.com.
[2] URL: http://www.hm.com.
[3] URL: https://pro.topsy.com.

Facebook (H&M)

Facebook wall data was extracted by a specialized big social data analytics tool called SODATO. SODATO[4] is an IT artifact, a software solution that is custom built for collecting, storing, processing, and analyzing big social data from social media platforms. The construction of SODATO is not only informed by our proposed theory of social data, but it is also methodologically built in adherence to Sein et al.'s (2011) Action Design Research (ADR) principles. Technically, SODATO utilizes the APIs provided by the social network vendors (e.g., Facebook open source API named as Graph API). Table 2 gives an overview of the social data collected by SODATO from the official Facebook walls of H&M.

Table 2: Overview of Facebook Data

Company	Official Facebook Wall: Name (id)	Time Period	Facebook Posts	Facebook Comments	Facebook Likes
H&M	Hm (21415640912)	January, 2009 → March, 2015	127,920	366,863	14,367,067

Sales (Apple and H&M)

Data for the Apple iPhone's quarterly sales in millions of units sold and H&M's quarterly revenues in billions of Swedish Kroner (SEK) were obtained from the respective companies' official annual reports. This concludes the presentation of the methods and tools used for data collection and overviews of the different big social datasets. We now discuss the third step in predictive analytics prescribed by Shmueli and Koppius (2011), data preparation.

Step 3: Data Preparation

Twitter (Apple: "iPhone")

We searched for the keyword "iPhone" in Topsy Pro, which then returned number of all tweets (i.e., Tweets, retweets, and replies) for the time period specified, and with sentiment numbers pre-calculated. These numbers form the basis for our prediction of one quarter sales of iPhones. We read the numbers of Tweets, and corresponding sentiment number in Topsy Pro on the screen, and inputted those numbers into Microsoft Excel. We employed calendar based quarters rather than the financial quarters of Apple for the modelling.

Facebook (H&M)

Facebook data was first fetched by SODATO via the Facebook Graph API and was then pre-processed and aggregated in order to make it available on demand for Analytics engine and at the end to the visualization module. The grouping of different analysis units was done in accordance

[4] URL: http://cssl.cbs.dk/software/sodato.

with the different attributes of the theory of social data (Social Graph: actors, actions and artefacts and Social Text: sentiment).

Sales (Apple and H&M)

As mentioned earlier, data for the Apple iPhone's quarterly sales in millions of units sold and H&M's quarterly revenues in billions of Swedish Kroner (SEK) were obtained from the respective companies' official annual reports. These were tabulated into Excel spreadsheets together with quarterly measures of social graph and social text.

Step 4: Exploratory Data Analysis

Shmueli and Koppius (2011) stated that during exploratory data analysis: "each question, rather than each construct, would be treated as an individual predictor. In addition to exploring each variable, examining the correlation table between BI and all of the predictors would help identify strong predictor candidates and information overlap between predictors (candidates for dimension reduction)" (p. 657).

Our objectives for the explorative data analytics were twofold: First, to build on the seminal regression model of Asur and Huberman (2010) for predicting movie revenues from twitter sales. Second, based on the Hierarchy of Effects (HoE) (Lavidge and Steiner 1961) and the AIDA (Attention, Desire, Interest, and Action) (Li and Leckenby 2007) domain-specific models of advertising and sales respectively, to explore different predictor variables, different data transformations of the predictor variables in terms of time lagging and different options for seasonal weighting of the predicted variable, sales.

We organize this section in the order of the two datasets (Apple iPhone tweets & H&M facebook) and describe the explorative data analysis conducted on the respective big social data sets that had already been collected and prepared.

"iPhone" Dataset

For the iPhone dataset, we selected the social data attributes of *social graph: actions* (tweets, re-tweets, replies and mentions) and *social text: keyword* ("iphone") and *social text: sentiments* (positive, negative, and neutral). We explored the temporal dynamics of the social data measures *social graph: actions* and *social text: sentiments* for the filter social text: keyword ("iPhone") directly on the Topsy Pro web site. We then explored the dataset by creating two predictor variables: quantity of tweets and quality of tweets as described below.

Quantity of Tweets

To provide an example, for the time period of September 10, 2013 to December 10, 2013, we made a data query in Topsy pro, specifying the period and searching for the phrase "iPhone" in all tweets (tweets, replies, retweets). For this example result was 44.62 million tweets and the corresponding sentiment number of 64.

Time Lagging of Tweets

As mentioned earlier, our predictive analytics method is informed by both the theory of social data and the AIDA and HoE domain-specific models. The key analytical challenge in social data predictive analytics is to model real-world outcomes from social data measures of social graph (actions, artefacts, activities and actions) and social text (topics, keywords, pronouns and sentiments). From the AIDA and the HoE domain-specific models and based on standard industry practice, we explored different options for time-lagging of social data measures as proxy for the sales funnel inherent in the time period between a potential customer becoming aware of the product, developing an interest in the product, having a desire for it and ultimately deciding to obtain it typically by a sales transaction. We experimented with different time-lags and found 20 days to be the statistically optimal value for the iPhone twitter dataset. As will be discussed later, we found different time lags for different datasets. It is important to note that even though the AIDA and HoE models can help in the exploration of the time lag in the first place and a partial explanation of its existence, they do not theoretically predict a particular value. This, we hope will be addressed with research advances in computational social science in general and predictive analytics in particular

Seasonal Weighting of Sales

Again, based on the AIDA and HoE models, and given the product life cycle of new models and new operating system releases of Apple iPhone, we conducted season weighting of the quarterly sales. Seasonal weights were calculated as the given quarter's proportion of the last calendar year. For example, the season weight for calendar Q3.2013 was calculated as below:

$$\frac{\text{Q3.2013 iPhone Sales}}{(\text{Q3.2013} + \text{Q2.2013} + \text{Q1.2013} + \text{Q4.2012})} = \frac{\text{33.8 million iPhone Sales}}{(33.80 + 31.24 + 37.43 + 47.79)} = \underline{0.225}$$

This proportion number 0.225 is then divided with 0.25 (0.225 / 0.25 = 0.90) to yield the season weight for that particular quarter. So the season weight for Q3.2013 is 0.90 which is multiplied with the 38.72 million tweets for that quarter.

Calculating season weights this way, always 4 quarters back in time, ensures that the calculation is always a mix of Q1, Q2, Q3 & Q4. So only one season weight has to be estimated, which is the latest number for prediction for next quarter. An estimated season weight for prediction must always go 1 year back. Next, we present the exploratory data analysis of the H&M dataset.

H&M Dataset: Following Shmueli and Koppius (2011)'s advice for exploratory data analysis step of predictive analytics, we explored the predictive power of several different variables constructed from the theory of social data. In summary, we created two categories of the social data attribute of *social graph: a*ctors (H&M and Non-H&M). We then calculated the distribution of the social data attribute of *social graph: artefacts* (posts, comments, and likes) across the two

actor types. With respect to the social data attribute of social text: sentiments (positive, negative, and neutral), based on the sentiment analysis of the social text artefacts (posts and comments) discussed earlier, we calculated distributions of sentiments across different kinds of artifacts and actors (i.e. positive sentiments on posts by H&M actors (wall administrators), positive sentiments on posts by Non-H&M actors etc.). We then calculated the quarterly aggregates of these different measures of social data attributes and evaluated the statistical correlation with respect to quarterly sales. Surprisingly, statistically significant positive correlations with quarterly revenues were observed for negative sentiments on total posts.

Logarithmic Transformation and Time lagging of Facebook Likes

Informed by the correlational analysis above and based on further exploratory data analysis with different predictor variables, we selected the logarithmic transformation of 40 days' time lagged total likes per quarter as the main predictor variable from the array of social data attributes listed in Tables 4 and 5 above.

Seasonal Weighting of Quarterly Sales

As with iPhone quarterly sales, we used a weighted measure of the quarterly revenues of H&M to account for seasonal variation of sales corresponding to fashion cycles (i.e., Fall, Winter, Spring and Summer Collections) and holidays across the different H&M markets.

Step 5: Choice of Variables

Choice of the predictor variables is based on careful considerations of theory, domain-specific knowledge and empirical association with predicted variables (Shmueli and Koppius, 2011). Based on exploratory data analysis, the following variables were chosen for the two big social datasets as summarized in Table 3.

Table 3: List of Chosen Predictor Variables

Company / Product	Time Period of Quarter	Seasonal Weighting of Dependent Variable [Sales]	Independent Variable #1 (including info on transformation)	Independent Variable #2	Time-Lagging of Independent Variable #1	Time-Lagging of Independent Variable #2
iPhone Sales (Quarterly)	Calendar Quarters	+	No of tweets over 3 months period	sentiment	20 days	20 days
H&M	Quarter ends 1 month before calendar quarter: Q4.2014 is from September 01 → November 30	+	LOG (No of total likes over 3 months period)	none	40 days	none

Step 6: Choice of Methods

As discussed earlier, we analytical objective was to not only build on but also extend the predictive modelling of Asur and Huberman (2010). As such, we chose regression modelling as the method and sought to extend the method by using time lagged and transformed predictor variables of social data measures and seasonally adjusted predicted variables.

Step 7: Evaluation, Validation and Model Selection

Our overall predictive analytics model for big social data analytics is stated below:

$$y = \beta_a \times A_t + \beta_p \times P_t + \beta_d \times D + \varepsilon$$

Where:

$A_t = \sum A_{st}$

A_{st} = Social media activity in terms of actions by actors on artifacts associated with sales at time t (Social Graph Attributes)

A_t = Accumulated time-lagged social media activity associated with sales at time t

P_t = Polarity at time t (Social Text Attribute)

D = Distribution factor (Sales Channel Attribute)

We now present the specific prediction models for the two big social datasets of iPhone & H&M.

Social Data Predictive Analytics Model for iPhone Sales

We modelled the relationship between iPhone sales and iPhone tweets in the period of 2010-2014 and excluded the period of 2007–2009. While the data for time period of 2007–2009 is noisy, the statistical association is relatively stable for 2010–2013 and gives an excellent correlation. Potential reasons could be historical growth of user base on Twitter, and also the development of socio-cultural practices of using twitter. The predictive model for iPhone sales is:

Predicted Sales of iPhones Sold (in millions) = WtweetRun * 0,6987228 + Sentiment * (-0,210626) + 22,845247 (intercept)

where

- WtweetRun is the season weighted tweets count for 3 months period time lagged by 20 days back from the sales quarter
- Sentiment is the sentiment for tweets for 3 month period time lagged by 20 days back the from sales quarter

Figure 3 presents the statistical output for the iPhone predictive model.

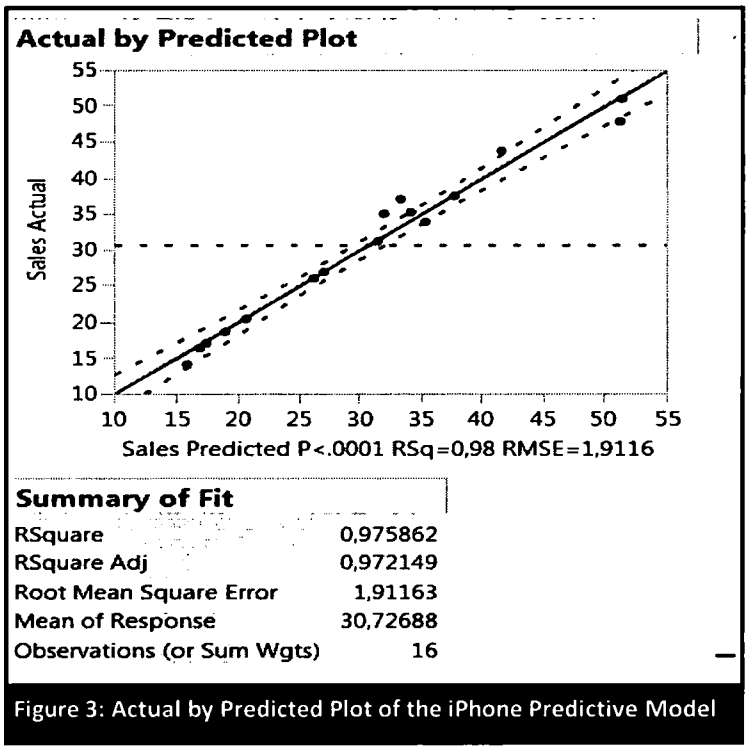


Figure 3: Actual by Predicted Plot of the iPhone Predictive Model

Figure 4 depicts the graph for the iPhone predictive model.

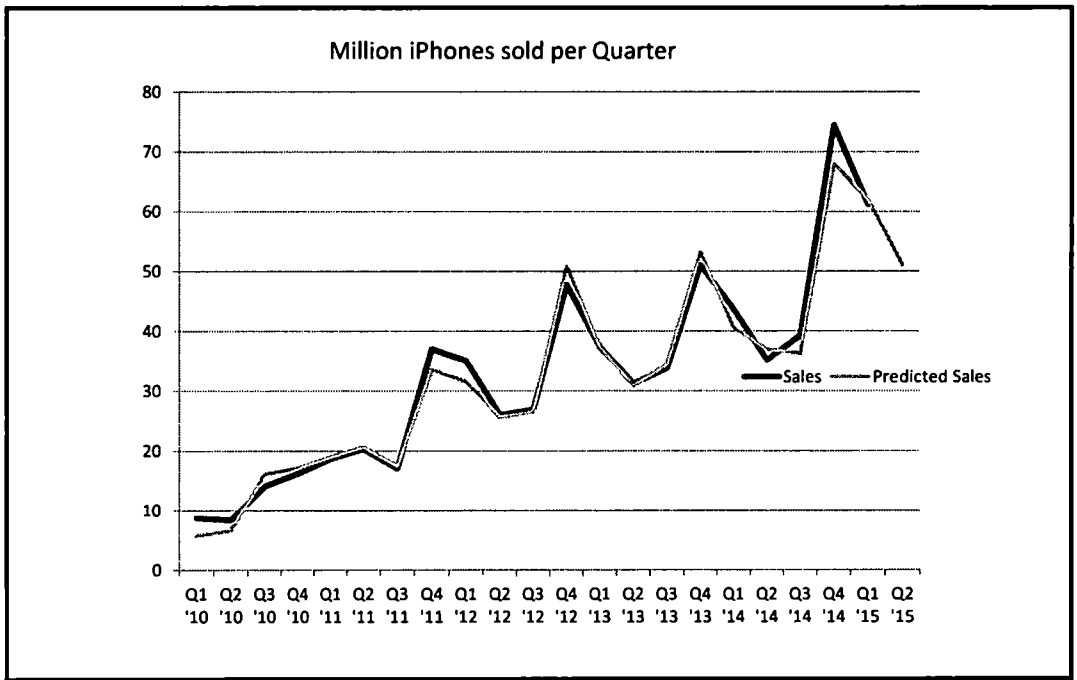


Figure 4: Predicted vs, Actual Sales of iPhones

Social Data Predictive Analytics Model for H&M Revenues

Based on the linear regression for the period 2011-2014, our predictive analytics model for 2014 is given by the following equation:
Predicted Revenue for H&M (in billions SEK) = 2,28 billion SEK * seasonweight * LOG (Facebook total likes time lagged by 40 days back over a 3 months period) + 5,45 billion SEK (the intercept)

Figure 5 presents the SAS output for the 2011-2014 predictive modelling

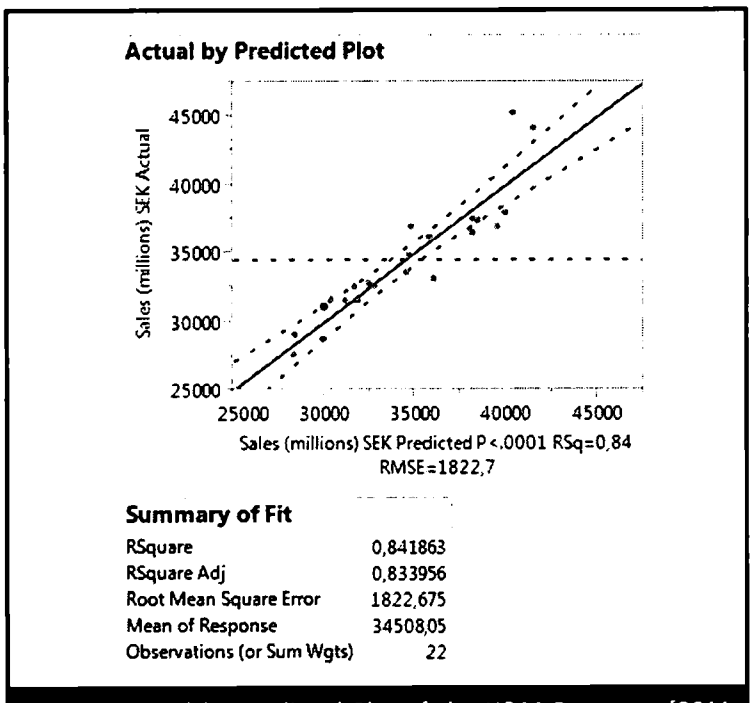


Figure 5: Actual by Predicted Plot of the H&M Revenues [2011-2014]

However, for the period of 2010-2013, based on the linear regression of data for 2009-2013, the predictive model is:

Predicted Revenue for H&M (in billions) = 1,67 billion SEK * seasonweight * LOG (facebook total likes time lagged by 40 days back over a 3 months period) + 13 billion SEK (the intercept)

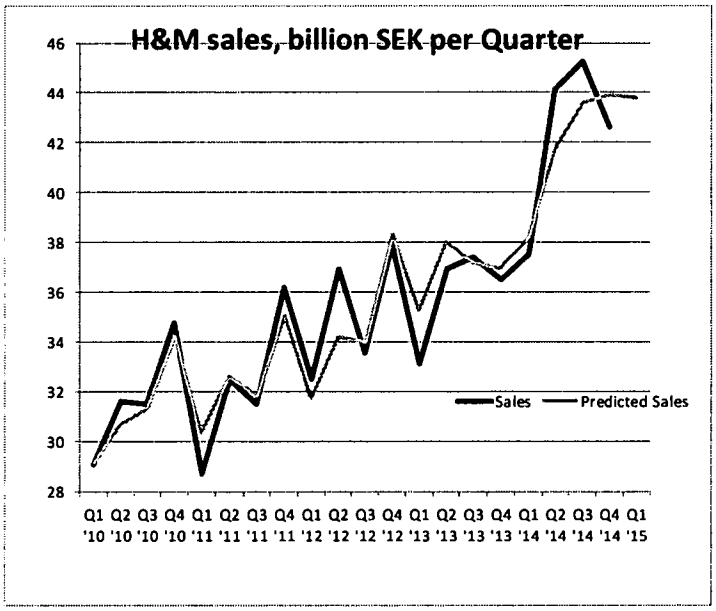


Figure 6: Predicted vs. Actual Revenues of H&M

Figure 6 depicts the combined chart of predicted vs. actual revenues of H&M

Step 8: Model Use and Reporting

In this step, we focus on predictive accuracy and meaning (Shmueli & Koppius, 2011). With regard to our prediction models, we observed a 5-10% average error from our prediction model with the actual sales data over 3 year period 2012-2014. In the case of iPhone, this average error is not far from the predictions of Morgan Stanley and IDC. For benchmarking purposes, we have identified a few leading prediction methods for iPhone sales.

- Morgan Stanley's "Alphawise Smartphone tracker" by Katy Huberty based on Google trend data, seasonal weighting, and socio economic factors[5].
- IDC's *Worldwide Quarterly Mobile Phone Tracker®, uses* bottom-up methodology[6]
- Steve Milunovich at UBS[7]

DISCUSSION

[5] URL: http://www.forbes.com/sites/chuckjones/2014/03/19/morgan-stanleys-alphawise-smartphone-tracker-has-iphone-demand-ahead-of-consensus.
[6] URL: http://www.idc.com/tracker/showproductinfo.jsp?prod_id=37.
[7] URL: http://www.forbes.com/sites/chuckjones/2013/12/03/ubs-analyst-milunovich-upgrades-apple-to-buy-with-650-price-target.

Though predictive analytics has been touted to be a major growth segment for research into social media, there is only a handful of studies to-date that have managed to capitalize on this opportunity. This paper thus takes a small but concrete step towards furthering this research agenda by advancing and validating a theory of social data for enhancing predictive analytics. Detailed implications for theory and practice are elaborated below.

Implications for Theory

This paper makes a novel contribution to extant literature on several fronts. First, past studies on social networks have typically progressed as two separate research streams with one seeking to comprehend the structural properties of such networks (i.e., social network analysis) (e.g., Johnson et al. 2014; Moser et al. 2013; Putzke et al. 2010; Shi et al. 2014; Trier 2008; Trier and Richter 2014; Whelan 2007; Whelan et al. 2013) and the other trying to infer value from the communicative content shared within these networks (i.e., sentiment analysis) (e.g., Cheung et al. 2012; Clemons et al. 2006; Jensen et al. 2013; Li and Hitt 2010; Mudambi and Schuff 2010). Yet, at the same time, there is evidence to suggest that invaluable insights could be gleaned from research that considers the structural properties and communicative content of social networks in tandem (see Butler et al. 2014; Chau and Xu 2012; Füller et al. 2014; Gasson and Waters 2013; Gray et al. 2011; Moser et al. 2013; Trier and Richter 2014). Therefore, in distinguishing between social graph and social text as constituent elements of social data, our proposed theory gives equal prominence to the two aforementioned research streams by embracing the structural properties and communicative content of social media.

Second, our theory of social data is the first to bring clarity to plausible dimensions that could be incorporated into empirical predictive models for social media (see Figure 1). By deriving constituent dimensions of social graph (i.e., actor, action, activity and artifact) and social text (i.e., topic, keywords, pronoun and sentiment), we enlarge the pool of options for applying predictive analytics to big social data. Third, we demonstrate the applicability of our proposed theory through the construction of empirical predictive models that are invariant to the kind of social media platform (i.e., Facebook and Twitter) from which data is extracted and the type of corporate entities (i.e., financial performance of H&M and iPhone) to be predicted, be it companies or products. In this sense, our proposed theory of social data can be deemed as a cornerstone for future studies of predictive big social data analytics to build upon.

Last but not least, beyond predictive analytics, we believe that our proposed theory of social data can also aid in the generation of holistic frameworks for computational social science in general and big social data analytics in particular. So far, computational methods, formal models and software tools for big social data analytics have been largely confined to graph theoretical approaches (Gross and Yellen 2005) in the likes of social network analysis (Borgatti et al. 2009),

which in turn is informed by the social philosophical approach of relational sociology (Emirbayer 1997). As far as we know, there is no other unified modeling approaches to social data that assimilates conceptual, formal, software, analytical and empirical domains (Mukkamala et al. 2013). Recent work (e.g., Vatrapu et al. 2014a, 2014b) has sought to outline an alternative approach to the predominant triad of relational sociology, graph theory and social network analysis, which are founded on associational sociology (Latour 2005), set theory and fuzzy set theory (Ragin 2000) as well as social set analysis (Mukkamala et al. 2014).

Implications for Practice

This paper should be of interest to practitioners for three reasons. First, our empirical results bear direct and indirect implications for companies. Naturally, a direct and obvious implication from this study is the proof that business performance can be predicted from big social data. By extracting and analyzing data from multiple social media platforms (i.e., Facebook and Twitter) to predict the financial performance of both companies (i.e., H&M) and products (i.e., iPhone), we are able to show that the predictive power of big social data is neither constrained by the social media platform nor the type of parameter to be predicted. For this reason, the indirect implications are that companies should proactively engage and strategically manage social media platforms in order to benefit from the strong correlations between social media interactions and sales performance. Second, by delineating social data into elements of social graph and social text, we provide companies with a schema of the elements to pay attention to on social media platforms. In order for companies to generate competitive advantage from social media, they must not only recognize the structural relationships within social networks, they must also value the opinions and sentiments embodied within social media content. Finally, this study is the first of its kind to take into account the existence of a time-lag from the moment a potential customer becomes aware of a product to the instance he/she decides to acquire it via a sales transaction when building empirical predictive models. In a way, this study highlights the importance of social media as an inexpensive forum for companies to continuously maintain product awareness in the minds of consumers.

Limitations

There are several limitations to the work reported here. First, we lack multiple cases to extensively evaluate and validate the overall prediction model. A second limitation is the emerging challenge for predictive analytics from social data associated with increasing sales in emerging markets such as China with its own unique social media ecosystem. By and large, the social media ecosystem of China does not overlap with that of Western countries to which Facebook and Twitter belong. We suspect that the effect of non-overlapping social media ecosystems might be somewhat ameliorated for Veblen goods such as iPhones given the

conspicuous consumption aspirations of a global middle class. This however remains an analytical challenge and restricts the predictive power of our H&M prediction model. A third limitation of the paper is that the theory of social data is limited to a cross-sectional framework of social data in terms of social graph (i.e., actors, actions, activities and artefacts) and social text (i.e., topics, keywords, pronouns and sentiments). As such, our theory of social data does not outline a process model, which might be more pertinent to predictive analytics. A fourth limitation arises from the representativeness of social media data. That said, as far as predictive analytics of real-world activities is concerned, social media datasets might be adequately representative as long as the basic premise of a social media action being a proxy for a user's attention to that particular real-world activity holds true. Our theory of social data will only cease to be valid if and when a user's social media action (such as a tweet about an "iPhone") is not a proxy for that user's attention towards the "iPhone" object. In our view, this fundamental disjunction between social media actions and real-world attention is the Achilles's Heel of predictive analytics with social data and might partially explain the spectacular drop in accuracy for once popular prediction models like the Google Flu Prediction System. A fifth and final limitation of our study, as far as our knowledge goes, is the lack of theoretical explanation for the empirical values for the time lags both in the nominal sense and the relative sense of divergence between Facebook and Twitter.

Future Work

For future work, we envision several projects that could spawn from this research as outlined below.

Going beyond the traditional and pre-dominant sentiment classification of social text and towards domain-specific classifiers such as AIDA and HoE for predicting sales. This will require not only sophisticated computational linguistics methods and tools but also critical contributions from domain experts (e.g., for training datasets in the case of supervised machine learning algorithms).

Investigating other predictor variables such as socio-economic factors, confidence, trust, loyalty etc. Essentially. Moving towards "thick models" of human users and narrowing the social media user and real-world consumer gap for non-digital products and services.

Combining social media data with other online sources such as Google Trends or in-house data of enterprise systems such as ERP and CRM.

REFERENCES

Asur, S. and Huberman, B. A. "Predicting the Future with Social Media," in Proceedings of the *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT) (Vol. 1), 2010, pp. 492-499.

Borgatti, S. P., Mehra, A., Brass, D. J. and Labianca, G. "Network Analysis in the Social Sciences," *Science* (323:5916), 2009, pp. 892-895.

Butler, B. S., Bateman, P. J., Gray, P. H. and Diamant, E. I. "An Attraction–Selection–Attrition Theory of Online Community Size and Resilience," *MIS Quarterly* (38:3), 2014, pp. 699-728.

Chau, M. and Xu, J. "Business Intelligence in Blogs: Understanding Consumer Interactions and Communities," *MIS Quarterly* (36:4), 2012, pp. 1189-1216.

Cheung, M. Y., Sia, C. L. and Kuan, K. K. "Is This Review Believable? A Study of Factors Affecting the Credibility of Online Consumer Reviews from an ELM Perspective," *Journal of the Association for Information Systems* (13:8), 2012, pp. 618-635.

Clemons, E. K., Gao, G. G. and Hitt, L. M. "When Online Reviews Meet Hyperdifferentiation: A Study of the Craft Beer Industry," *Journal of Management Information Systems* (23:2), 2006, pp. 149-171.

Emirbayer, M. "Manifesto for a Relational Sociology," *The American Journal of Sociology* (103:2), 1997, pp. 281-317.

Füller, J., Hutter, K., Hautz, J. and Matzler, K. "User Roles and Contributions in Innovation-Contest Communities," *Journal of Management Information Systems* (31:1), 2014, pp. 273-308.

Gasson, S. and Waters, J. "Using A Grounded Theory Approach to Study Online Collaboration Behaviors," *European Journal of Information Systems* (22:1), 2013, pp. 95-118.

Gray, P. H., Parise, S. and Iyer, B. "Innovation Impacts of Using Social Bookmarking Systems," *MIS Quarterly* (35:3), 2011, pp. 629-643.

Gross, J. L. and Yellen, J. *Graph Theory and Its Applications*, CRC press, 2005.

Hussain, A. and Vatrapu, R. "Social Data Analytics Tool," DESRIST 2014, *Lecture Notes in Computer Science* (LNCS), 8463(Springer), 2014, pp. 368–372.

Jensen, M. L., Averbeck, J. M., Zhang, Z. and Wright, K. B. "Credibility of Anonymous Online Product Reviews: A Language Expectancy Perspective," *Journal of Management Information Systems* (30:1), 2013, pp. 293-324.

Johnson, S. L., Faraj, S. and Kudaravalli, S. "Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment," *MIS Quarterly* (38:3), 2014, pp. 795-808.

Lassen, N., Madsen, R. and Vatrapu, R. "Predicting iPhone Sales from iPhone Tweets," in Proceedings of the *18th IEEE Enterprise Computing Conference* (EDOC 2014), Ulm, Germany, 2014.

Latour, Bruno (2005). *Reassembling the social an introduction to actor-network-theory*. Oxford New York:

Oxford University Press. ISBN 9780199256044.

Lavidge, R. J. and Steiner, G. A. "A Model for Predictive Measurements of Advertising Effectiveness," *Journal of Marketing* (25:6), 1961, pp. 59-62.

Li, H. and Leckenby, J. "Examining the Effectiveness of Internet Advertising Formats," in D. Schumann & E. Thorson (eds.), *Internet Advertising: Theory and Research*, Lawrence Erlbaum Associates, 2007, pp. 203-224.

Li, X. and Hitt, L. M. "Price Effects In Online Product Reviews: An Analytical Model And Empirical Analysis," *MIS Quarterly* (34:4), 2010, pp. 809-831.

Moser, C., Ganley, D. and Groenewegen, P. "Communicative Genres as Organizing Structures in Online

Communities–of Team Players and Storytellers," *Information Systems Journal* (23:6), 2013, pp. 551-567.

Mudambi, S. M. and Schuff, D. "What Makes A Helpful Online Review? A Study Of Customer Reviews on Amazon.Com," *MIS Quarterly* (34:1), 2010, pp. 185-200.

Mukkamala, R., Hussain, A. and Vatrapu, R. "Towards a Formal Model of Social Data," *IT University Technical Report Series*, TR-2013-169, 2013. [Available online at: https://pure.itu.dk/ws/files/54477234/ITU_TR_54472013_54477169.pdf, accessed October 14, 2014]

Mukkamala, R., Hussain, A. and Vatrapu, R. "Towards a Set Theoretical Approach to Big Data Analytics," in *Proceedings of IEEE Big Data 2014*, Anchorage, United States of America, 2014.

Putzke, J., Fischbach, K., Schoder, D. and Gloor, P. A. "The Evolution Of Interaction Networks In Massively Multiplayer Online Games. *Journal of the Association for Information Systems* (11:2), 2010, pp. 69-94.

Ragin, C. C. *Fuzzy-Set Social Science*, University of Chicago Press, 2000.

Sein, M., Henfridsson, O., Purao, S., Rossi, M. and Lindgren, R. "Action Design Research," *MIS Quarterly* (35:1), 2011, pp. 37-56.

Shi, Z., Rui, H. and Whinston, A. B. "Content Sharing in A Social Broadcasting Environment: Evidence From Twitter," *MIS Quarterly* (38:1), 2014, pp. 123-142.

Shmueli, G. and Koppius, O. R. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), 2011, pp. 553-572.

Trier, M. "Research Note-Towards Dynamic Visualization For Understanding Evolution Of Digital Communication Networks," *Information Systems Research* (19:3), 2008, pp. 335-350.

Trier, M. and Richter, A. "The Deep Structure of Organizational Online Networking–An Actor□Oriented Case Study," *Information Systems Journal* (Advance copy) 2014.

Vatrapu, R. "Understanding Social Business," In K. B. Akhilesh (ed.), *Emerging Dimensions of Technology Management*, New Delhi: Springer, 2013, pp. 147-158.

Vatrapu, R. and Suthers, D. "Intra-and Inter-Cultural Usability in Computer-Supported Collaboration," *Journal of Usability Studies* (5:4), 2010, pp. 172-197.

Vatrapu, R. Cultural Considerations in Computer Supported Collaborative Learning," *Research and Practice in Technology Enhanced Learning* (3:2), 2008, pp. 159-201.

Vatrapu, R. K. "Explaining Culture: An Outline of a Theory of Socio-Technical Interactions," in *Proceedings of the 3rd International Conference on Intercultural Collaboration*, Copenhagen, Denmark, 2010, pp. 111-120.

Vatrapu, R., Mukkamala, R. R. and Hussain, A. "A Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Findings," in *Computational Social Science: Contagion, Collective Behaviour, and Networks*, Oxford: University of Oxford, 2014a, pp. 22-24. [available online at: http://cssworkshop.oii.ox.ac.uk]

Vatrapu, R., Mukkamala, R. R. and Hussain, A. "Towards a Set Theoretical Approach to Big Social Data Analytics: Concepts, Methods, Tools, and Empirical Findings," in *Proceedings of the 5th Annual Social Media & Society International Conference 2014*, Toronto, Canada, 2014b.

Whelan, E. "Exploring Knowledge Exchange In Electronic Networks of Practice," *Journal of Information Technology* (22:1), 2007, pp. 5-12.

Whelan, E., Golden, W. and Donnellan, B. "Digitizing The R&D Social Network: Revisiting The Technological Gatekeeper," *Information Systems Journal* (23:3), 2013, pp. 197-218.

Anvendelse af hedonisk boligprisindeks

Jakob Holmgaard[1], Lars Hervig Jacobsen[2]

[1]: Priser og Forbrug, Økonomisk Afdeling, Danmarks Statistik

[2]: Priser og Forbrug, Økonomisk Afdeling, Danmarks Statistik

Resumé

Dette papir giver en introduktion til anvendelsen af hedoniske boligprisindeks i Danmarks Statistik. Forskellige hedoniske metoder præsenteres samt deres implementering i praksis. Ligeledes præsenteres nogle foreløbige resultater for den eksisterende Ejendomssalgsstatistik.

Baggrund

I forbindelse med udvikling af et nyt prisindeks for nybyggede boliger har Danmarks Statistik valgt at tage udgangspunkt i hedonisk regression, hvilket ikke tidligere har været anvendt i den løbende produktion. Hedonisk regression er ikke kun velegnet til boligprisindeks, men kan også anvendes indenfor andre områder, herunder bilprisindeks. Udviklingen af hedonisk baserede prisindeks involverer både regressionsteknikker og konstruktion af kvalitetskorrigerede prisindeks. I dette papir fokuseres på det sidstnævnte.

Boliger som sælges på forskellige tidspunkter er forskellige med hensyn til beliggenhed, størrelse og standard og er derved ikke direkte sammenlignede, da kvaliteten af boligerne er forskellig. For bedre at nå frem til den rene prisændring mellem boliger handlet i to perioder kan man anvende forskellige indeksmetoder, der justerer for kvalitetsforskelle over de to perioder. Der findes groft sagt fire indeksmetoder til at korrigere for problematikken omkring "konstant kvalitet"[1]. I parentes er angivet hvilken statistik/producent, der anvender metoden:

1. Kvadratmeterpris-metode (Boligmarkedsstatistikken)
2. Gentagne salg-metoden (Boligøkonomisk Videncenter)
3. Sales Price Appraisal Ratio-metoden (SPAR) (Ejendomssalgsstatistikken, Danmarks Statistik)
4. Hedonisk regression (under udvikling for nybyggede boliger i Danmarks Statistik)

[1] Se Eurostat(2013).

I Danmarks Statistik anvendes SPAR-metoden i Ejendomssalgsstatistikken, nærmere bestemt den værdivægtede SPAR-metode. Ved denne metode indekseres afstandsprocenten, som er forholdet mellem summen af salgspriserne og summen af de tilhørende ejendomsvurderinger i en given periode:

$$P_{0:t} = \frac{afspct^t}{afspct^0} \tag{1}$$

, hvor

$$\text{afspct}^t = \frac{\frac{\sum_{i=1}^{n} P_i}{n}}{\frac{\sum_{i=1}^{n} V_i}{n}} = \frac{\sum_{i=1}^{n} P_i}{\sum_{i=1}^{n} V_i} \tag{2}$$

SPAR-metoden kan ikke anvendes for *nybyggede boliger*, da man ikke har samhørende par af salgspriser og vurderinger. I stedet har Danmarks Statistik valgt at indsamle prisoplysninger og bygningskarakteristika for udvalgte byggevirksomheder og anvende hedonisk regression til at beregne prisindeks. Denne metode anvendes desuden af statistikbureauerne i Norge, Sverige og Nederlandene.

Indenfor traditionel indeksteori defineres et værdiindeks således:

$$V_{0:t} = P_{0:t}^{LA} \cdot Q_{0:t}^{PA} = P_{0:t}^{PA} \cdot Q_{0:t}^{LA} \tag{3}$$

Der gælder således, at værdiindekset ($V_{0:t}$) kan findes som produktet af Laspeyres prisindeks ($P_{0:t}^{LA}$) og Paasche mængdeindeks ($Q_{0:t}^{PA}$) eller som produktet af Paasche prisindeks ($P_{0:t}^{PA}$) og Laspeyres mængdeindeks ($Q_{0:t}^{LA}$), jf. dualitetssætningen [Ulstrup Johansen, Per og Trier, Mikael (2010)][2]. De to prisindeks defineres således:

$$P_{0:t}^{LA} = \frac{\sum_{i=1}^{n} P_i^t \cdot Q_i^0}{\sum_{i=1}^{n} P_i^0 \cdot Q_i^0} \quad \text{(Laspeyres prisindeks)} \tag{4}$$

$$P_{0:t}^{PA} = \frac{\sum_{i=1}^{n} P_i^t \cdot Q_i^t}{\sum_{i=1}^{n} P_i^0 \cdot Q_i^t} \quad \text{(Paasche prisindeks)} \tag{5}$$

Til det formål at korrigere for værdien af kvalitetsændringer vil vi bruge en regressionsmodel til at beskrive udvalgte karakteristikas påvirkning af boligernes salgspris. Nærmere bestemt formuleres en ligning, hvor en boligs logaritmiske salgspris er en funktion af karakteristika ved ejendommen (f.eks. boligareal, beliggenhed, alder etc.) beskrevet ved K kovariate:

$$\ln(\mathrm{p}^i) = \beta_0^i + \sum_{k=1}^{K} \beta_k^i \cdot \mathrm{X}_k^i + \varepsilon^i \tag{6}$$

[2] Vi har her bevidst udeladt Fisher prisindeks og Fisher mængdeindeks.

Hvor det antages at forventningsværdien af støjledet er lig nul, $E[\varepsilon^i] = 0$. Konventionen om den log-lineære model er motiveret af ønsket om at stabilisere variansen og dermed reducere effekten af heteroskedasticitet samt ønsket om at beregne Jevon indeks (geometriske indeks). Ligeledes kan de kovariate log-transformeres efter behov. Parametrene kan estimeres ved mindste-kvadraters-metode (OLS), hvilket giver anledning til at opskrive et udtryk for den dekomponerede log-transformerede boligsalgspris, udtrykt ved gennemsnitsværdierne for karakteristika ($\overline{X}$) og parameterestimaterne ($\hat{\beta}$):

$$\overline{\ln(p^t)} = \hat{\beta}_0^i + \sum_{k=1}^{K} \hat{\beta}_k^i \cdot \overline{X}_k^i \qquad (7)$$

Bevægelsen over tid i karakteristika repræsenterer det kvalitative og dermed mængdemæssige element i gennemsnitsprisens udvikling. Bevægelsen over tid i β-parametrene, repræsenterer det prismæssige element. Dette illustrerer, at gennemsnittet af de rå salgspriser $\exp\left(\overline{\ln(p^t)}\right)$ kan opfattes som en værdi, der på basis af den hedoniske metode kan dekomponeres i en pris og mængde.

Hedoniske prisindeksmetoder

Der findes fire typer af hedoniske metoder til at beregne hedoniske prisindeks. Triplett(2006) inddeler disse metoder i to overordnet grupper: Den direkte metode og den indirekte metode[3]:

1. Tidsdummy metoden
2. Karakteristika metoden

} direkte metode

3. Pris-imputerings metoden
4. Re-pricing metoden

} indirekte metode

Ved den direkte metode tages der udgangspunkt i at man kan opskrive et værdiindeks som produktet af et prisindeks og et mængdeindeks[4]. Man gør her brug af *gennemsnitsværdier*, hvorfor der ikke behøves at være det samme antal observationer i prisindeksformlen for de to perioder, der sammenholdes. Ved den indirekte metode gør man brug af prissæt for de enkelte produkter, hvilket minder om traditionel indeksteori. Der skal dermed være det samme antal observationer i prisindeksformlen for de to perioder, der sammenholdes, hvilket sikres ved at tage udgangspunkt i et antal repræsentantvarer, som følges over tid. Ved re-pricing metoden forudsættes det, at man på forhånd har et datasæt med ukorrigerede prissæt for begge perioder, der

[3] Metoderne findes rundt om i litteraturen med forskellige navne. Karakteristika-metoden kaldes i Hjort-Andersen(1986) for *skyggeprismetoden*. Re-pricing metoden kaldes i Triplett(2006) for *quality adjustment method*.

[4] Jf. den føromtalte dualitetssætning.

indgår i prisindeksformlen. Metoden vil så korrigere disse prissæt, så de afspejler den rene prisudvikling. Ved pris-imputerings metoden anvender man produkterne for den ene periode som repræsentantvarer og anvender de imputerede priser for netop disse produkter i den anden periode, hvorfor der altid vil være det samme antal observationer i de to perioder *efter* imputering på trods af, at der ikke nødvendigvis findes lige mange observationer i det oprindelige datasæt. Hvis den samme datakilde anvendes i prisindeksformlen og i den hedoniske regression, så svarer pris-imputerings metoden, re-pricing metoden og karakteristika-metoden til hinanden. Vi ser bort fra de indirekte metoder og vil i det efterfølgende kort præsentere karakteristika-metoden og tidsdummy-metoden.

Karakteristika-metoden:

Ligning (4) og (7) giver anledning til at definere et hedonisk Laspeyres-prisindeks ved:

$$P_{0:t}^{LA} = \frac{exp(\widehat{\beta}_0^t + \sum_{k=1}^{K} \widehat{\beta}_k^t \cdot \bar{X}_k^0)}{exp(\widehat{\beta}_0^0 + \sum_{k=1}^{K} \widehat{\beta}_k^0 \cdot \bar{X}_k^0)} \quad \text{(hedonisk Laspeyres prisindeks)} \tag{8}$$

Hvor gennemsnitsværdierne for variablene, beregnet for basisperioden, indgår i tæller og nævner. Hvis man i stedet anvender de beregnede gennemsnit for den aktuelle periode fremkommer et hedonisk Paasche-prisindeks, jævnfør ligning (5):

$$P_{0:t}^{PA} = \frac{exp(\widehat{\beta}_0^t + \sum_{k=1}^{K} \widehat{\beta}_k^t \cdot \bar{X}_k^t)}{exp(\widehat{\beta}_0^0 + \sum_{k=1}^{K} \widehat{\beta}_k^0 \cdot \bar{X}_k^t)} \quad \text{(hedonisk Paasche prisindeks)} \tag{9}$$

Hvor parameterestimaterne findes ved at udføre regression på data fra hhv. den aktuelle periode og basisperioden. Bemærk at SPAR metoden er et specialtilfælde af karakteristika metoden, idet afstandsprocenten kan estimeres ved at opskrive ligning (6) med kun én variabel (vurdering) uden konstantled, ved hjælp af pseudo poisson maximum likelihood estimatoren i stedet for OLS [Ramalho, Esmeralda A. (sept. 2011)].

Tidsdummy-metoden:

Ved denne metode introduceres en tidsdummy-variabel, D^i, der giver anledning til, at estimere parameteren, γ^i, der antages, at beskrive prisudviklingen over tid, alt andet lige. Således at salgsprisen bestemmes ved:

$$\ln(\mathrm{p}) = \beta_0 + \gamma^i \cdot \mathrm{D}^i + \sum_{k=1}^{K} \beta_k \cdot \mathrm{X}_k + \varepsilon \tag{10}$$

Hvor, $\mathrm{D}^t = 1$, i den aktuelle periode, mens, $\mathrm{D}^0 = 0$, i basisperioden. Prisindekset kan da bestemmes direkte ved:

$$P_{0:t}^{TD} = exp(\gamma^i) \tag{11}$$

Hvor, γ^i, er estimeret ved at regressere på pooled data for periode 0 og t. Herved har man antaget, at parameterestimater er konstant over de to tidsperioder og der dermed ikke forekommer nogen kvalitetsændring over dette tidsrum. Tidsdummy-metoden kan derfor opfattes som resultatet af at lægge restriktioner på karakteristika-metoden.

Implementering af metoderne:

Underliggende for den hedoniske metode er et omfattende arbejde med databehandling, modelselektion og modeldiagnose, der skal sikre, at man dels har valid data for de tilgængelige variable i datasættet, at man anvender den mest sandsynlige model givet data og at den udvalgte model overholder de grundlæggende OLS antagelser. Endvidere kan man undersøge robustheden af parameterestimaterne ved konventionelle krydsvalideringsmetoder, eller hvis data er tilgængelige for adskillige tidsperioder er det ligeledes muligt, at betragte hvorvidt parameterestimaterne fluktuerer over tid og om de skifter fortegn. Vi vil ikke beskrive disse procedurer i deltaljer her, men blot nævne, at de er en forudsætning for udarbejdelsen af hedoniske prisindeks, jævnfør best practice.

Det er grundlæggende for metoden, at tidsperioderne defineres, hvoraf den aktuelle periode følger af frekvensen af det indeks man ønsker at publicere. Ligeledes følger tidsperioden af referenceperioden for tidsdummy-metoden af den aktuelle periode, da man blot pooler data for de to perioder der har samme tidsinterval. referenceperioden for karakteristika metoden er mere arbitrær. Man kan eksempelvis vælge en referenceperiode på ét eller to år, til trods for, at den aktuelle periode er på ét kvartal. I dette tilfælde er det nødvendigt at rense ud for prisudviklingen *indenfor* referenceperioden, hvilket her bliver gjort ved at introducere tidsdummy-variable for hvert kvartal [Statistics Norway (2012)]. Tidsdummy-variablene i referenceperioden korrigerer de beta-estimater, som der skal bruges i prisindeksformlen, men de inkluderes ikke selv i prisindeksformlen. En anden mulighed er, at estimere parametrene for hvert kvartal i basisperioden og beregne et gennemsnit for disse [Statistics Netherlands (July 5th 2013)].

Skift af referenceperioden

Ved karakteristika-metoden vælges typisk en fast referenceperiode, for eksempelvis det seneste kalenderår. Da vi ikke ønsker, at skift af referenceperioden i sig selv påvirker prisindekset, skal det kædede prisindeks beregnes efter følgende formel:

$$I_t^{kædet} = I_t^{ny\ ref} \cdot \frac{I_{t-1}^{gml\ ref}}{I_{t-1}^{ny\ ref}} \quad \text{(ved skift af referenceperiode)} \qquad (12)$$

I de perioder hvor der ikke skiftes referenceperiode, da sættes faktoren $I_{t-1}^{gml\,ref} = 1$. dvs:

$$I_t^{kædet} = \frac{I_t^{ny\,ref}}{I_{t-1}^{ny\,ref}} \quad \text{(ved fast referenceperiode)} \tag{13}$$

Bemærk, at ved tidsdummy-metoden beregnes det kædede prisindeks ved at kumulere den kvalitetskorrigerede prisudvikling fra periode t-1 til t, jf. ligning (11):

$$I_t^{kædet} = exp(\gamma^i) \cdot I_{t-1}^{kædet} \quad \text{(tidsdummy-metoden)} \tag{14}$$

Ud/indskiftning af kovariate i regressionsmodellen

Jævnfør traditionel indeksteori, så skal et produkt indgå i mindst to perioder før man kan beregne en prisudvikling. Overføres denne tankegang til hedonisk regression, så skal de samme egenskaber (kovariate) indgå i begge de perioder, der sammenholdes. Vi anbefaler derfor, at ud/indskiftning af kovariate kun bør ske i forbindelse med skift af referenceperioden, hvilket foretages én gang årligt ved karakteristika metoden (i.e. modelselektionen foretages på baggrund af referenceperioden). Såfremt en variabel viser sig at være insignifikant i den aktuelle periode, så skal den stadig medtages i beregningerne medmindre man ekstraordinært også fjerner den fra referenceperioden.

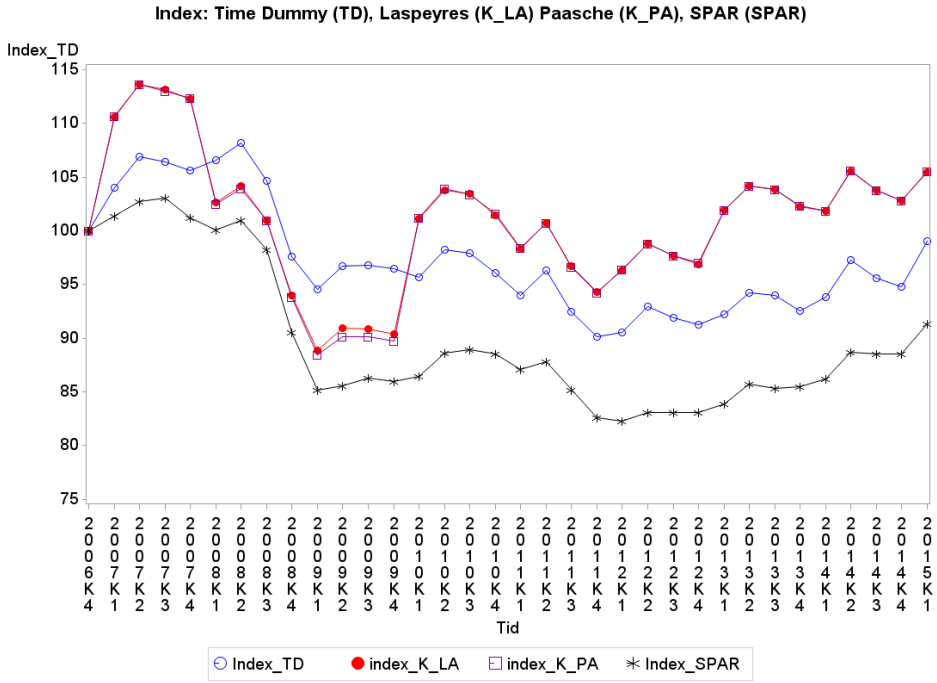
Tilgængelige data

Datagrundlaget er her dannet på baggrund af udtrak fra Bygnings- og Boligregistret (BBR), samt enkelte data fra [Boliga.dk, 2015]. Herved er der indsamlet data for de nævnte variable i Tabel 1. Bemærk at vi har valgt at kode antal værelser, antal badeværelser og boligens alder, som kategoriske dummy-variable da sammenhængen mellem salgsprisen og de her nævnte forklarende variable fremstår som ikke-lineær. Endvidere har vi inkluderet to, mere eller mindre, opfindsomme dummy-variable, udfra om ordet ”strand” indgår i vejnavnet og om boligen befinder sig i et dyrt område, jævnfør statistikken ” Top 100 dyreste veje i hele landet” fra [Boliga.dk, 2015]. Dette er motiveret af hypotesen om, at det er dyrere at bo ved kystlinien, samt at snob-effekten puster prisen på boligerne op. Dette er et forsøg på, at indkludere variable der beskriver boligens beliggenhed, om end mangelfuldt.

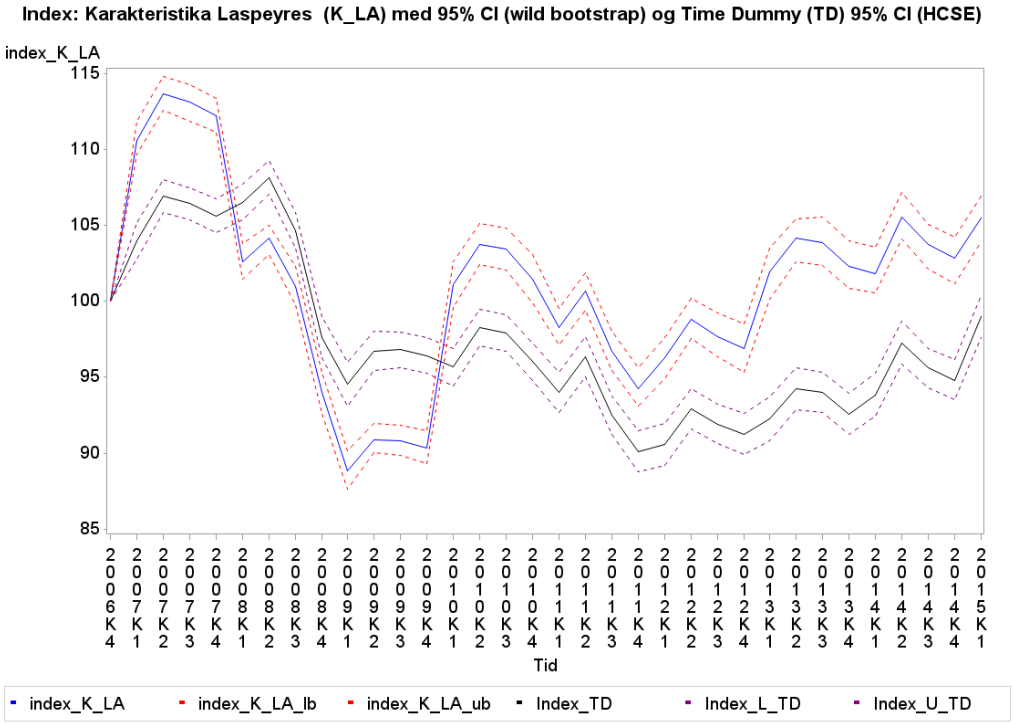
Beskrivelse	Variabel	Min	Max	Gns	Std afv
Salgspris:	ln_salgspris	11,51 (0.1 mio kr)	17,034 (25 mio. kr)	14,22	0.64
Husaeral	areal_hus	25	687	135,85	40,23
Antal værelser:					
1-3 stk	vaerelser_1	0	1	0,168	0,374
4 stk	vaerelser_2	0	1	0,349	0,477
5-7 stk	vaerelser_3	0	1	0,454	0,498
8-10 stk	vaerelser_4	0	1	0,029	0,167
Antal badeværelser:					
1 stk	badevaerelser_1	0	1	0,722	0,448
2 stk	badevaerelser_2	0	1	0,268	0,443
3 stk	badevaerelser_3	0	1	0,009	0,094
4 stk	badevaerelser_4	0	1	0,0003	0,0176
Varmeinstallation:					
Fjernvarme/blokvarme (radiatorsystemer el. varmluftanlæg)	varme_1	0	1	0,508	0,500
Centralvarme fra eget anlæg, et-kammer fyr	varme_2	0	1	0,402	0,490
Ovne (kakkelovne, kamin, brændeovne o.l.)	varme_3	0	1	0,007	0,083
Varmepumpe	varme_4	0	1	0,022	0,146
Elovne, elpaneler	varme_5	0	1	0,060	0,238
Andet	Varme_99	0	1	0,0005	0,022
Alder:					
0-15 år	alder_1	0	1	0,096	0,295
16-34 år	alder_2	0	1	0,174	0,379
34+ år	alder_3	0	1	0,730	0,444
Landsdele:					
København by	zone_1	0	1	0,03	0,171
København omegn	zone_2	0	1	0,078	0,268
Nordsjælland	zone_3	0	1	0,091	0,287
Bornholm	zone_4	0	1	0,014	0,116
Østsjælland	zone_5	0	1	0,047	0,212
Vest- og sydsjælland	zone_6	0	1	0,118	0,322
Fyn	zone_7	0	1	0,098	0,297
Sydsjælland	zone_8	0	1	0,149	0,356
Østjylland	zone_9	0	1	0,154	0,361
Vestjylland	zone_10	0	1	0,096	0,294
Nordjylland	zone_11	0	1	0,126	0,332
Diverse dummy:					
Top 100 dyrest vejnavne (Boliga.dk)	dummy_luksus	0	1	0,003	0,052
”Strand” indgår i vejnavnet.	dummy_strand	0	1	0,012	0,109

Resultater

Vi præsenterer her den estimerede prisudvikling for enfamiliehuse i perioden 1. kvartal 2007 til 1. kvartal 2015. De udvalgte hedonsike modeller er en undergruppe af den fulde model beskrevet ved de kovariate i Tabel 1. Det er for omstændigt at præsentere resultatet af modelselektionen for samtlige perioder her, hvilket dog vil være en del af dokumentationen for statistikproduktionen. I stedet viser vi blot det endelige resultat i form af det beregnede prisindeks ved de fire metoder tidligere beskrevet, hhv. SPAR-metoden (ligning(1)), Laspeyres prisindeks ved karakteristika-metoden (ligning (8), Paasche prisindeks ved karakteristika-metoden (ligning (9)) og tidsdummy-metoden (ligning (11)), (se figur 1). Ligeledes vises resultaterne for hhv. Laspeyres prisindeks ved karakteristika-metoden og tidsdummy-metoden med estimerede 95 pct. usikkerhedsintervaller i separat plot (se figur 2).



Figur 1



Figur 2

Diskussion

Af de her præsenterede prisindeks, forekommer karakteristika metoden som værende den mest anvendelige. Antagelsen om uændrede karakteristika over to tidsperioder vil ofte være for restriktiv, hvilket begrænser anvendeligheden af Tidsdummy-metoden. Karakteristika metoden har flere frihedgrader, om end de kovariate er begrænset til de udvalgte i referenceperioden.

En af fordelene ved den hedoniske tilgang er, at man kan drage nytte af al tilgængelig data. Det kan imidlertid diskuteres om man har tilstrækkelig med data til at designe en fornuftig model. Eksempelvis er det tvivlsomt, om hvorvidt man kan modellere en boligs salgspris uden at tage højde for beliggenheden, hvilket kun er gjort i begrænset omfang her. Endvidere er det mangelfuldt, at der ikke er data der beskriver om boligen har været igennem en omfattende renovering, så som nyt tag, køkken etc.. Hertil må man have in mente, at den udvalgte model optimalt kun repræsentere et lokalt optimum i et uendeligt parameterrum. Modellen kan i princippet altid forbedres, og derfor stiller det krav til tilstrækkelig dokumentation ved udgivelse af statistikken (retninglinier der ikke er fulgt i denne artikel), således at brugerne kan foretage sin egen vurdering, samt få indsigt i de valg der er foretaget og mulighed for at reproducere resultatet og eventuelt forbedre det. Her ligger dog muligvis en af metodens svagheder begravet, idet det kan være svært at kommunikere de tekniske detaljer videre til en lægmand.

En betydelig del af arbejdsprocessen består i at klargøre data til analyse. Her er det nødvendigt at fejlsøge, samt opstille en metode for håndtering af manglende observationer. Det er givet, at man enten bør fjerne eller rette fejlbehæftede observationer. Bemærk at detekterede outliers forbliver i data medmindre det kan verificeres at der er tale om en fejlbehæftet observation. Non-respons skal håndteres med en vis påpasselighed, idet en ukritisk udrensning kan medføre en bias. Endvidere er det en optimeringsprocedure, der består i valget mellem et stort udvalg af kovariate versus antallet af tilgængelige observationer, idet flere variable øger sandsynligheden for tilstedeværelsen af én manglende værdi for en given observation.

I det tilfælde at data stammer fra en stikprøve, kan man tage højde for de vægtede observationer ved at anvende vægtet-mindste-kvardraters metode (WLS) i stedet for OLS. Dette er imidlertid ikke anbefalet, da det er problematisk, at anvende stikprøvevægte der stammer fra en lav-dimensionel stratifikation (Typisk 2-3 dimensioner), når den hedoniske model ofte har andre og flere dimensioner. Herved risikerer man, at lade én observation vægte for adskillige observationer, som den ikke deler egenskaber med og dermed får man et mindre korrekt estimatet ved WLS end med OLS. En anden ulempe ved WLS er at det kan fremprovokere heteroskedasticitet. Derfor kan det være en bedre løsning at foretage en uvægtet OLS estimering på hvert enkelt stratum hvorved et samlet prisindeks kan beregnes ud fra en vægtet sum af de estimerede prisindeks i strata [se Eurostat(2013)].

Sidst, men ikke mindst, så skal de her præsenterede resultater ikke ses om endelige, da metoden stadigvæk er under udvikling. Dermed vil der gå nogle arbejdstimer før metoden bliver en fast del af statiskstikproduktionen.

Referencer:

1. CENEX håndbog (2009): Statistics and Science. Handbook on the application of quality adjustment methods in the Harmonised Index of Consumer Prices, Vol. 13, Federal Statistical Office of Germany. Developed within the European project "CENEX HICP Quality Adjustment".
2. Eurostat (2013): Handbook on Residential Property Prices Indices (RPPIs)
3. Hjort-Andersen, Christian (1986): Hedoniske regressioner: Hvad koster en meter bil? Nationaløkonomisk Tidsskrift, Bind 124, s. 90-105.
4. Statistics Norway (2012): Boligprisindeksen, dokumentasjon af metode.
5. Ramalho, Esmeralda A. (sept. 2011): Hedonic functions, hedonic methods, estimation methods and Dutot and Jevons house price indexes: are there any links?
6. Statistics Netherlands (July 5th 2013): Method description. New dwellings; output price indices building costs, 2010=100.
7. Ulstrup Johansen, Per og Trier, Mikael (2010): Praktisk statistisk metode for økonomer, 3. udgave.
8. Triplett, Jack (2006): Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes
9. http://www.boliga.dk/statistik/20130227.aspx?list=2

Modeling Advertising Effects in a Multi-media Environment - A Latent Class Latent Markov Chain Approach

Carsten Stig Poulsen, professor, Ph.D.

Pål Børresen, M.Sc., Schibsted Media Group

Corresponding author:

Carsten Stig Poulsen,

Strandboulevarden 114,

DK-2100 Copenhagen, Denmark.

carsten.stig.poulsen@gmail.com

A new approach to the measurement of advertising effects in a multi-media environment, based on consumer panel data, is presented. We formulate a dynamic, individual model that can be used for descriptive, predictive, segmentation, and - potentially - optimization purposes. It may even be utilized during an on-going campaign to monitor and eventually correct the media plan. The model is based on recent advances in latent Markov modeling and views advertising and its effects as interactive stochastic processes that unfold over time, influenced by the impact of advertising. The model is demonstrated by a real, but disguised application.

Keywords: advertising, effects, multi-media, latent Markov model, segmentation

The Challenge

The measurement of advertising and its effects on marketing goals has occupied marketing practitioners and researchers for decades. This should come as no surprise as practitioners have a legitimate interest in documenting the returns on their advertising investments, and research firms and scientists have used their skills and creativity to come up with tools and models that might fill the gap.

In his survey of fifty years of cross-media research (Assael 2010) identified two developments:

- from one medium at a time (silo) to two or more media in interaction (synergy)
- from opportunity-to-see (OTS) to opportunity-to-act (OTA)

He concludes, however, that synergy has yet to achieve its full potential. He writes: "Of most importance is the lack of reliable measures of cross-media effects. Ideally, single-source systems would measure multi-media exposure and purchase behavior for the same respondent. [...] Until adequate measures of interactive media effects are developed, cross-media effects will not reach its full potential", (op. cit., p. 42).

In this paper we propose a modeling approach that we claim might fill this gap. The plan of the paper is as follows: First, to fix ideas we present the advertising campaign that will be used to demonstrate the approach. Then the conceptual model of "how advertising works" is presented with an overview of the proposed modeling framework. The various submodels and the ideas underlying the approach are presented, keeping the mathematics at a minimum. (The interested reader can acquire a more technical version of the paper by contacting the corresponding author.) The results of applying the model to data from the campaign are covered. The paper concludes with a summary and some future extensions of the approach.

The Application: Introducing a new potato chip

The advertising campaign

In the spring of 2011 a Norwegian producer of snacks introduced a new potato chip with a rifled surface that potentially increases the spicy taste. The introduction was accompanied by an advertising campaign, involving two media channels, TV and the Web. In addition, promotional activities like in-store displays were initiated even before the start of the advertising campaign. Hence, some consumer awareness of the new chip at the start of the campaign is to be expected.

The media plan

As the proposed model is designed to measure the effects of a specific advertising campaign an important input to the model is the media plan, i.e. the choice of media, their impacts and timing e.g. in terms of GRPs or insertions, cf. figure 1. As can be

seen it involves TV and the Web (banner ads) during the weeks 8 till 13. The plan was designed prior to the decision by the company to act as a case illustration of the model.

FIGURE 1

The Media Plan

	TV Media		Net media									
	TV Ch. 1	*TV Ch. 2*	*Site 1*			*Site 2*			*...*	*Site 10*		
Week	Days with showings	Days with showings	Days with banner ads	Omega	Effective banner days	Days with banner ads	Omega	Effective banner days		Days with banner ads	Omega	Effective banner days
7	0	0	0	100%	0	0	100%	0		0	100%	0
8	1	1	0	100%	0	0	100%	0		0	100%	0
9	7	7	2	100%	2	0	100%	0		7	100%	7
10	7	7	1	100%	1	0	100%	0		7	100%	7
11	0	0	0	100%	0	0	100%	0		0	100%	0
12	7	7	0	100%	0	0	100%	0		0	100%	0
13	1	1	0	100%	0	0	100%	0		0	100%	0
14	0	0	0	100%	0	0	100%	0		0	100%	0
15	0	0	0	100%	0	0	100%	0		0	100%	0

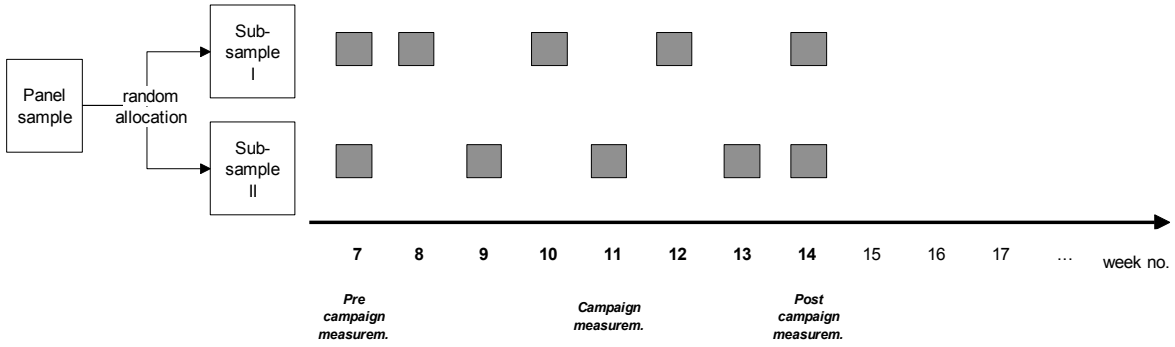
Setting up the panel

A panel of consumers in the target group (15 – 59 years old) was set up by recruiting respondents in an already existing internet panel. Data was to be collected weekly starting one week prior to the campaign (week 7) and ending one week after (week 14). By including a pre- and a post-measurement we obtain an increase in the variability of the media impacts, making the effects of the campaign more detectable.

In order to reduce costs, respondent fatigue and memory effects of the previous answers only half of the panel members were asked at each wave. This was done by splitting the entire sample into two random halves at the outset of the measurement, cf. figure 2.

FIGURE 2

Study Design



The measurements
In each of the eight panel waves several sets of questions were posed:

- Eating habits about chips or snacks
- Awareness, perceptions, preferences, and purchases for a selected set of four major brands
- Recall of advertising for the selected brands in TV, newspapers, and on the internet
- Usage of media during the previous week

The questions used for illustration in this paper are given table 1.

TABLE 1

The questions used in the case

Question	Response categories
Q1. Seen ad in TV last week for Our Brand	Yes/no/DK
Q2. Our Brand has introduced a new chip with rifled surface	Yes/no/DK
Q3. Intention to buy Our Brand at next purchase occasion	Yes/no/DK
Q4. How often did you watch [list of two TV channels] at home or elsewhere during last week?	0,1,…,7 days out of 7/DK
Q5. How often did you visit [list of 10 websites] during last week?	0,1,…,7 days out of 7/DK

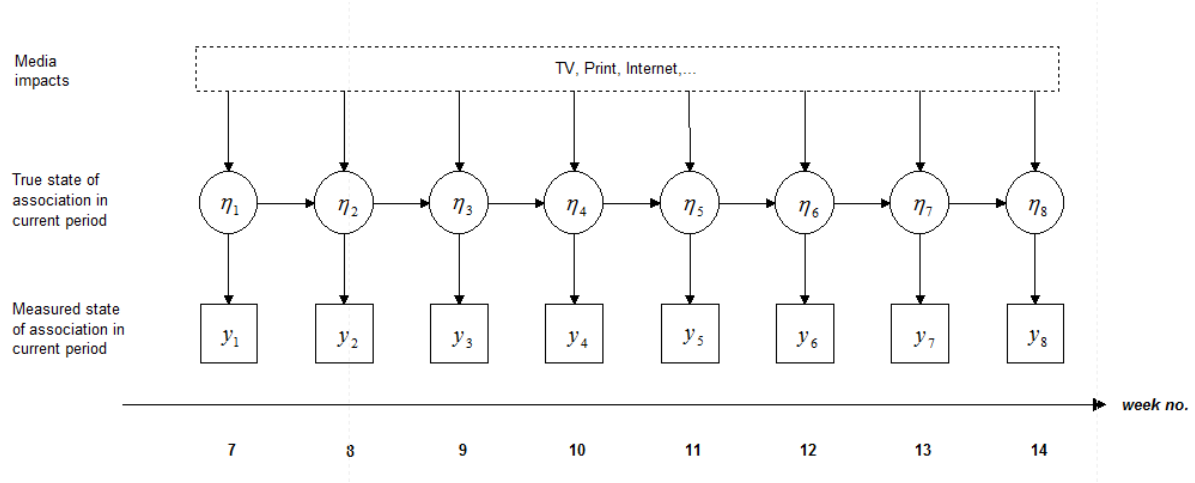
Conceptual Foundation of the Effects Model

Advertising and its effects on the target group are seen as stochastic processes that develop over time, involving latent or unobserved true states that are recursively interdependent. To be specific, one of the KPIs of the described campaign is that people should associate the chips of Our Brand with a 'rifled' surface. Exposure to a TV commercials or banner advertisements on the Web is assumed to work by moving over time the individual from her initial state of 'no association', to a state of 'association'. However, neither of these states, 'no association' nor 'association', are directly observable. The state related to this KPI is measured by Q2 in table 1, but when responding to questions like this, people make errors. To account for this, we introduce the distinction between the manifest measure (the item used for questioning) and the latent state, representing the true, error-free state of the individual. Further, advertising is seen as working at the latent state rather than manifest level, and by eliminating or controlling for measurement errors, we stand a better chance to detect

the true effects. This is the prime justification for working with processes involving latent states. Figure 3 gives a general picture of the process and its measurements.

FIGURE 3

A path-diagrammatic presentation of the model with one latent KPI and errors in measurement



Readers familiar with LISREL, (Joreskog and Van Thillo 1972), will recognize the diagrammatic and notational conventions. Manifest measured variables are depicted by squares, circles represent latent variables. However, the η-variables are discrete, latent variables, representing a stochastic process $(\eta_t = 0,1; t = 0,1,\text{K}\ ,T)$, where 0 is the latent state 'no association' and 1 is latent state 'association'. The arrows between the η-variables indicate dependencies over time between consecutive points and the process is said to be of first-order. The arrows between the latent η-variables and the manifest y-variables signify that y_t is an indicator of η_t, but conditionally independent of all other η- and y- variables. The general statistical framework for analyzing these kinds of data is latent Markov, (Langeheine and Van de Pol 1990) or latent transition modeling, (Collins and Lanza 2010).

The Proposed Modeling Framework

Figure 4 provides an overview of the model.

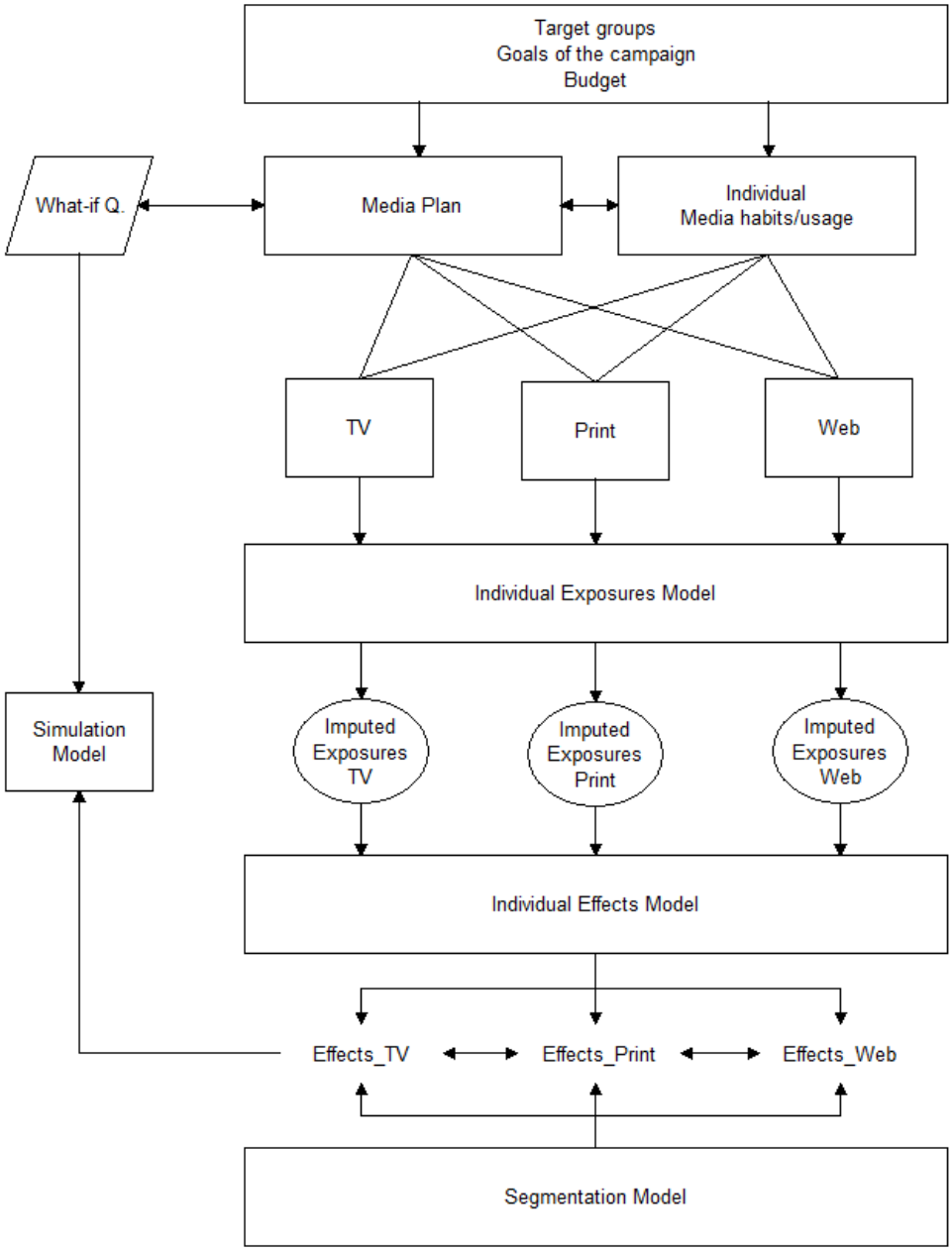
The model requires four types of input and contains four submodels.
The inputs are:

- Definition of target groups, if the market is seen as segmented
- Statement of the goals of the campaign, precise enough to allow the goals to be expressed as Key Performance Indicators (KPIs)

- The media plan containing the employed media, their impacts in terms of number of insertions and timing, and
- The media usage or habits of the individual respondents in the panel.

FIGURE 4

Graphical representation of the modeling approach



The four submodels each have their specific purpose:

- The *Exposure model* aims at predicting individual exposures to advertising, based on the insertions from the media plan and stated usage of each medium. Usage can be habits (invariant during the campaign) or weekly usage, allowing for dynamic changes during the campaign.

- The *Effects model* is the core of the approach. It uses a latent or hidden Markov chain model with the estimated individual exposures acting as covariates for explaining and predicting the (transitions between) states of the Markov model, defined in terms of one or more of the KPIs of the campaign. The effects are coefficients in logistic regression models that allow for main as well as interaction effects of the media.

- The important issue of heterogeneity in consumer response is taken care of in the *Segmentation model*. It may take three venues:
  1. Firstly, prior defined target groups can be compared in terms of KPI traces assuming common effects coefficients, but possibly differences in media usage.
  2. Secondly, the effects model may be re-estimated, conditional on each target group. Problems of sparse data may restrict this approach.
  3. Thirdly, a latent class approach may be employed, allowing for unobservable segments with coefficient heterogeneity. This approach does not required prior defined target groups, but points to the existence of segments that may be profiled in terms of other variables in the data set, besides media habits.

- Finally, given an estimated model, whether segmented or unsegmented, changes in the media plan can be tracked in the *Simulation model* as time traces of the KPIs. This provides a managerially more interpretable overview of the consequences of various media plans, compared with inspecting the coefficients of the logistic model. In addition, under certain conditions of stationarity the effects of the campaign beyond the campaign period can be forecasted.

Application to the case

We are now ready to apply the model to the data of the potato chips case. For this purpose we use the software Latent Gold in its syntax version 4.5, (Vermunt and Magidson 2008). The method of estimation is maximum likelihood (ML). Among the virtues of the program we mention that it handles the high-dimensional problem of computing the latent state probability vector by using a version of the Baum-Welch algorithm, (Vermunt 2003). In addition, the ML estimation handles missing values, due to design or non-response, under the assumption of missing at random (MAR), (Rubin 1976). To compare the goodness-of-fit of the models we use the information theoretical measure BIC, (Carlin and Louis 2009).

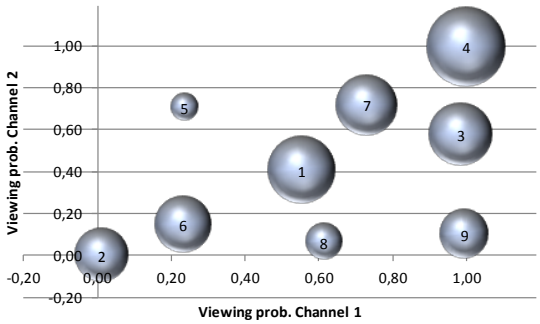
Applying the Exposure model

The media plan involves two media groups, TV with two Channels 1 and 2, and banner ads on 10 website on the Web. We treat each media group as independent in usage. Within each group the channels/websites are analyzed simultaneously.

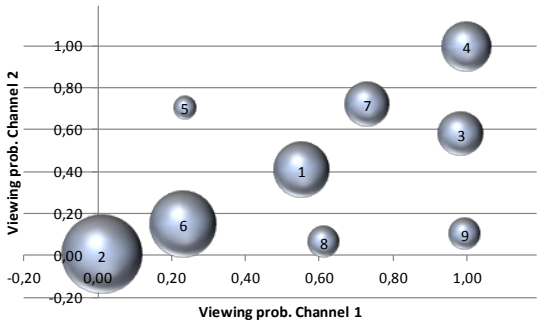
Also here, we apply a latent Markov modeling approach. Due to space limitation we focus on the TV media group and its two channels. Based on BIC an unrestricted latent Markov model with 9 classes (states) proved to give a reasonable fit and figure 5 provides a graphical display of the estimated structure and the dynamics, when we let the process reach 'steady state'.

FIGURE 5

Initial frequency states
for Channel 1 and Channel 2
Area proportional to size



Frequency states
for Channel 1 and Channel 2 in 'steady state'
Area proportional to size



From figure 5 we conclude that Channel 1 has the largest audience, since most of the states have a higher viewing probability for this channel compared to Channel 2. Only state 5 comprising 3% of the respondents exhibit a preference for Channel 2. Furthermore, when looking at the dynamics of the state proportions developing in accordance with the estimated transition matrix there is a clear tendency towards less viewing on both channels. Due to this development average exposures can be expected to decrease over time, for any fixed media impact.

Although interesting by itself, this structural analysis of the viewing pattern of the two TV channels, a more important result in this context is, that it can be utilized to compute individual, time-dependent viewing probabilities, based on Bayes' theorem. Combined with the insertions of TV advertising according to the media plan, cf. table 1, expected number of exposures are computed and assigned to individuals.

The same approach is used to compute individual exposures, generated by banner advertisement on the 10 websites, included in the media plan, and we now have covariates for the TV as well as the Web-impact to be included in the Effects model.

Applying the Effects model with a single KPI

We want to measure the effects, if any, of exposures to advertising in TV and on the Web on the KPI "association of Our Brand to rifled surface", cf. table 1. We shall do this by fitting alternative specifications of our basic latent Markov chain model with one process, and the imputed, individual exposures to TV commercials and banner ads, including possible interaction, as covariates. We shall use a stepwise approach, starting with a naïve null-model, the hypothesis of (marginal) independence between indicators, as baseline, and compare the subsequent, more substantive models in terms of BIC, with this.

TABLE 2

Overview of various fitted models for the one KPI process model

Model no. and description	BIC	No. of parameters
1. Null model: Independence between manifest indicators; no measurement errors; no covariates	4340,2819	1
2. Allowance for state dependency (carry-over); no measurement errors (manifest Markov chain); no covariates	3250,3671	3
3. Allowance for state dependency (carry-over) and measurement errors (latent Markov chain); no covariates	3154,3785	5
4. Allowance for state dependency (carry-over) and measurement errors (latent Markov chain); TV and web exposures included as covariates. Main effects and interaction effect.	3155,3515	8
5. Allowance for state dependency (carry-over) and measurement errors (latent Markov chain); TV and web exposures included as covariates. Main effects only.	3148,5966	7
6. Allowance for state dependency (carry-over) and measurement errors (latent Markov chain); TV and web exposures included as covariates. Full response surface model specified.	3162,6780	10
7. Allowance for state dependency (carry-over) and measurement errors (latent Markov chain); TV and web exposures included as covariates. Main effects interact with lagged latent state.	3159,9657	9

Model no. and description	BIC	No. of parameters
8. Allowance for state dependency (carry-over) and measurement errors (latent Markov chain); TV exposures included as covariate. Main effect and squared term included.	3143,2684	7
9. Allowance for state dependency (carry-over) and measurement errors (latent Markov chain); TV exposures included as covariate. Main effect only.	3141,9575	6

The single parameter of the null model (model 1) is the common response probability of the indicator over time.

Model 2 allows for dependency across time by introducing a first-order, carry-over effect for transitions between states, i.e. a (manifest) Markov chain, and we see that this gives a much better model as measured by the decrease in BIC. It implies that being in a given state e.g. "no association to rifled surface" is informative about an individual's likely state at the next time point. As we shall see, chances are that the individual remains in the *same* state, but that can strictly speaking only be concluded *after* inspecting the estimated transition matrix. Two additional parameters, the transition probabilities, are estimated.

In model 3 we extend the manifest Markov model to include state uncertainty in the form of errors in the response variables, i.e. we introduce a latent Markov model. Again, two additional parameters, the conditional response probabilities, given the latent state, are needed, but it seems to pay off as the BIC-measure decreases further.

At this point we have reached a stationary, first-order latent Markov chain as the best model to describe the data. We now introduce the two covariates for media impacts,

$$z_{t-1;TV} \text{ and } z_{t-1;Web},$$

indicating the (lagged) exposures to TV and the Web, imputed by the Exposure model. Since these covariates are individual and time-varying, we are in fact facing a heterogeneous, non-stationary latent Markov model. Starting with model 4 that includes the two main effects as well as one interaction effect of the two media it is seen from the increase in BIC that the three additional effects parameters are not justified. However, since this model is at the core of our modeling effort, we shall display the estimated model in full and provide some interpretation.

Although all parts of the model, the reliability matrix, $\mathbf{\Delta}^{\eta}$, the initial state distribution $\boldsymbol{\pi}_0^{\eta}$, and the transition matrix $\mathbf{T}^{\eta}$ can be cast as logit-models, the interpretation of the

results is more intuitive, when expressed in probabilities. It is only the introduction of covariates in $\mathbf{T}^{\eta}$ that makes interpretation somewhat more difficult. We therefore provide the estimated model $\hat{\boldsymbol{\Delta}}^{\eta}, \hat{\boldsymbol{\pi}}_0^{\eta}$ and $\hat{\mathbf{T}}_0^{\eta}$ with

$$z_{t-1;TV} = z_{t-1;Web} = 0.$$

Subsequently, we provide the estimated logit-model for $\hat{\mathbf{T}}^{\eta}$ with covariates included.

TABLE 3

Estimated parameters of model 4 in table 2

$$\hat{\boldsymbol{\Delta}}^{\eta} = \begin{bmatrix} 0{,}9611 & 0{,}0389 \\ 0{,}0649 & 0{,}9351 \end{bmatrix} \qquad \hat{\boldsymbol{\pi}}_0^{\eta} = \begin{bmatrix} 0{,}8920 & 0{,}1080 \end{bmatrix} \qquad \hat{\mathbf{T}}_0^{\eta} = \begin{bmatrix} 0{,}9647 & 0{,}0353 \\ 0{,}0388 & 0{,}9612 \end{bmatrix}$$

From $\hat{\boldsymbol{\Delta}}^{\eta}$ we see that a close, but not perfect association between the true, latent state and the measured response exists. The matrix is to be read horizontally: given the latent state, e.g. "truly associate brand with rifled surface", the probability of responding "yes" is 0,9351, while the negative "no" with conditional response probability of 0,9611 is somewhat more reliable. As we know from model 3 compared to model 2, assuming perfect measurements (1.0 on the main diagonal) gives a poorer model. The initial distribution $\hat{\boldsymbol{\pi}}_0^{\eta}$ indicates that 11% of the respondents start in true state "associates…". Basically, the goal of the campaign is to increase this percentage by moving individuals from the "no association" to the "association" state. $\hat{\mathbf{T}}_0^{\eta}$ provides insight into the *intrinsic* dynamics of this process, if left to itself with no advertising. Hence, the trajectory of the proportion of individuals in the "association" state over time, based on $\hat{\mathbf{T}}_0^{\eta}$, may serve as the dynamic benchmark when the effects of advertising are evaluated.

We now include the effects of the covariates and provide the estimated logit-model for the transition matrix with main and interaction effects. Note, that when presenting the results, the focus is on state $\eta_t = 1$ while $\eta_{t-1} = 0$ is reference state.

The estimated logit equations for the initial state probability and the transition probabilities with covariates included is displayed in (1)-(2) and graphically presented in figure 6.

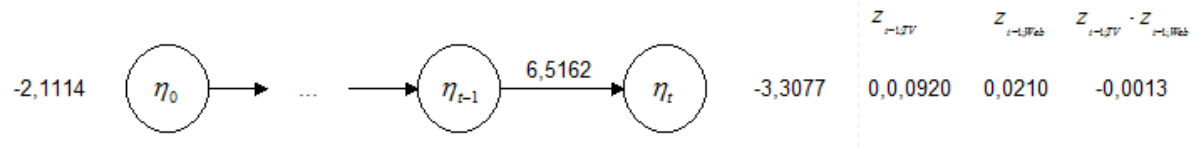
$$\text{logit}\left(\pi_{\eta_0}\right) = -2{,}1114 \tag{1}$$

$$\text{logit}\left(\pi_{\eta_t|\eta_{t-1};z_{t-1;TV},z_{t-1;Web}}\right) = -3{,}3077 + 6{,}5162 \cdot \eta_{t-1} + 0{,}0920 \cdot z_{t-1;TV} + 0{,}0210 \cdot z_{t-1;Web} - 0{,}0013 \cdot z_{t-1;TV} \cdot z_{t-1;Web}$$

(2)

FIGURE 6

Graphical illustration of the estimated logit model



The negative constant -2,1114 indicates that initially the marginal probability of being in state "no association" is much larger that state "association" (in fact, we already know this distribution from table 3), while -3,3077 is the logit for 'association' when the previous state is 'no association', $\eta_{t-1} = 0$, and no advertising takes place,

$$z_{t-1;TV} = z_{t-1;Web} = 0 .$$

The largest, positive coefficient is the carry-over effect from the previous state, η_{t-1}. This underlines the importance of including this effect in the model, and it indicates the size of the challenge for the advertising to overcome this. The interpretation is that the logit of being in state "association" in the next period is 6,5162 larger, if the individual is in state "association" in the current period than if he is in state "no association". As can be seen from the coefficient of the covariates, they do have positive, albeit modest, main effects. An additional exposure will increase the logit for $\eta_t = 1$with 0,0920 for TV and 0,0210 for the Web. Also note the negative coefficient -0,0013 of the interaction term. However, neither the main effect of Web-advertising nor the coefficient of the product term is statistically significant, using BIC as well as asymptotic $t-$tests.

To illustrate the flexibility of the approach we fit some additional models. In model 5 the product term of model 4 is excluded, and the BIC is reduced. Model 6 specifies a full response surface for the two media, i.e. squared as well as cross-product terms are included. The increase in the number of estimated parameters is not justified as measured by BIC. Model 7 tests the hypothesis that the (main) effects are dependent on (interact with) the lagged state variable η_{t-1}. This seems not to be the case. Finally, we test model 8 with only TV as the medium, but with squared exposures included to test for decreasing returns of advertising. The estimated model is:

$$\text{logit}\left(\pi_{\eta_t|\eta_{t-1};z_{t-1;TV}}\right) = -3,4797 + 6,5829 \cdot \eta_{t-1} + 0,2569 \cdot z_{t-1;TV} - 0,0125 \cdot z^2_{t-1;TV}$$

The coefficient to the squared term is significant at the .02-level. However, the BIC-measure points to model 9 with only main effects for TV included.

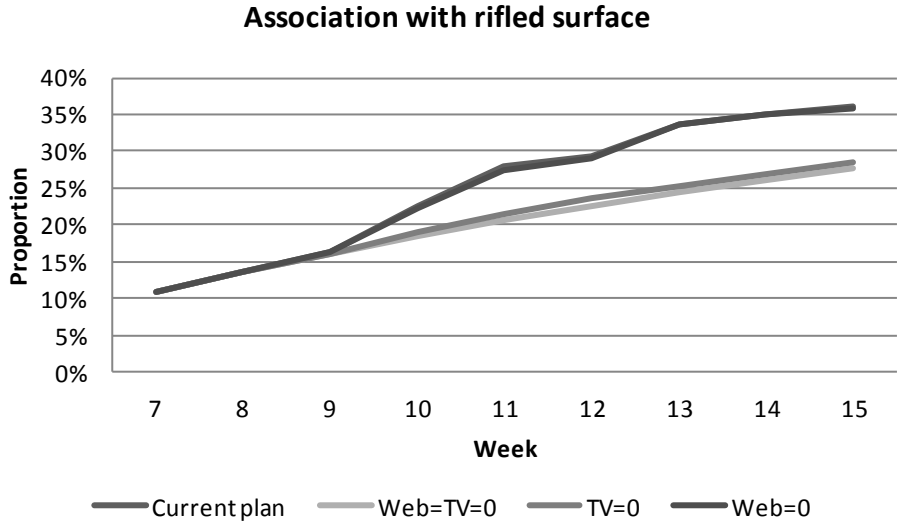
On the effects of advertising

The model provides a kind of answer to the question that seems to dominate the discussion of the effects of advertising in a multimedia environment, see e.g. (Naik and Raman 2003). Are the media working separately (main effects), together (interaction) and if so, with positive or negative synergy? It should be clear, however, that the conclusions are totally dependent on how the model that is used is formulated mathematically. The present modeling approach is (generalized) linear in parameters and predictors when the dependent variable is the logit, and the estimated coefficients to the predictors, including product terms, have a clear interpretation as partial effects *ceteris paribus*. However, replacing the logit with the odds makes the model multiplicative in predictors, and interaction is built-in as part of the formulation. The parameters may now be interpreted as elasticities, but the effect of one medium depends on the levels of all other.

The question of how media work, partially and jointly, may be answered in a managerially more meaningful way within the modeling framework by simulating the effect of alternative media plans on the KPI in question. To illustrate, using model 4 we compare the dynamic development in the proportion of individuals in state "association" in four scenarios: no advertising $KPI_t^{0,0}$, TV alone $KPI_t^{TV,0}$, Web alone $KPI_t^{0,Web}$, and TV and Web jointly $KPI_t^{TV,Web}$, all based on the actual media plan, cf. figure 7. $KPI_t^{0,0}$ represents the intrinsic development of the proportion over the campaign period, due to factors outside the model. Hence, this may serve as the dynamic bench-mark for evaluating the effects of the two media, alone or together.

FIGURE 7

Development in KPI "Association with rifled surface" in four different advertising scenarios



It is clear from figure 7 that TV advertising is the most important contributor the increase in KPI. A measure of the efficacy of TV and Web advertising might be the entire cumulated increase over time in KPI, attributable to the media impact. Since advertising works beyond the campaign period, these effects long-run should be included when a campaign is evaluated. This problem is addressed by using the path of the latent KPI process towards the steady-state as the bench-mark. Therefore, we propose as the measure of efficacy for medium m_1:

$$\psi_{m_1} = \frac{\sum_{t=0}^{\infty}\left(KPI_t^{m_1,0} - KPI_t^{0,0}\right)}{\sum_{t=0}^{T} GRP_t^{m_1}} \tag{3}$$

$$\psi_{m_1,m_2} = \frac{\sum_{t=0}^{\infty}\left(KPI_t^{m_1,m_2} - KPI_t^{0,0}\right)}{\sum_{t=0}^{T}\left(GRP_t^{m_1} + GRP_t^{m_2}\right)} \tag{4}$$

An economic evaluation of the campaign (ROI) and subsequent optimization requires increases in KPIs to be expressed in monetary terms (assuming the costs of the campaign are known), and that is easier if the KPI is more closely related to behavior, e.g. purchase intention or brand choice. Therefore, models that include more than one KPI, e.g. in the form of a 'hierarchy-of-effects', is of major interest. The approach presented here can be extended to such models as well. Space limitations restrict us from going into that here.

Conclusions

To sum up, the framework for analyzing the effects of advertising presented in this paper:

- Reflects the *specific goals* of the advertising campaign
- Uses the media plan as input and *simulates* the consequences of changes
- Introduces a *novel approach* to estimating average exposures in a multi-media environment using a 'wafers' approach rather than a 'silo' approach,
- *Is dynamic*, i.e. is able to describe, to explain, and to predict the effects of advertising as a phenomenon that extends over time,
- Reflects *state-dependency*, i.e. the effects of advertising may depend on the state of the respondent in terms of previous exposures, previous state of response, etc.
- Incorporates *response uncertainty/errors of measurement* in the indicators (questions) that are used in the survey
- Is capable of modeling *main* and *interaction effects* of multiple media
- Allows for *non-stationarity* and *individuality* through the inclusion of individual time-constant and time varying *covariates* in the state processes

- Models *recursive systems* of processes, e.g. *hierarchy-of-effects* hypotheses by defining each stage in the hierarchy as separate, but related processes with restrictions on transitions across the levels of the hierarchy
- Models individual *response heterogeneity* as a group analysis with *prior* defined groups or *unobserved* heterogeneity by introducing a latent segmentation variable

Still, further developments are needed, e.g.

- Other types of input data: count data, banner click, unobtrusive exposure data, choice modeling, etc. Especially, the new possibilities of tracking individual behavior on the internet seem promising and should be incorporable seamlessly into the framework
- Media planning optimization and ROI-considerations also await to be fully developed.

References

Assael, Henry (2010), "From Silos to Synergy," Journal of Advertising Research, 51, 42-58.

Carlin, Bradley P. and Thomas A. Louis (2009), Bayes and Empirical Bayes Methods for Data Analysis (3rd ed.). Boca Raton, FL: Chapman & Hall.

Collins, Linda M. and Stephanie T. Lanza (2010), Latent Class and Latent Transition Analysis. Hobroken, N.J.: Wiley & Sons, Inc.

Joreskog, Karl G. and M. Van Thillo (1972), "LISREL: A General Computer Program for Estimating a Linear Structural Equation System Involving Multiple Indicators of Unmeasured Variables." Latest version 8.8 Jan 2012 ed. Princeton, N.J.: Educational Testing Service.

Langeheine, Rolf and Frank Van de Pol (1990), "A Unifying Framework for Markov Modeling in Discrete Space and Discrete Time," Sociological Methods Research, 18 (4 (May)), 416-41.

Naik, Prasad A. and Kalyan Raman (2003), "Understanding the Impact of Synergy in Multimedia Communications," Journal of Marketing Research (JMR), 40 (4), 375-88.

Rubin, Donald B. (1976), "Inference and missing data," Biometrika, 63, 581-92.

Vermunt, J.K. and J. Magidson (2008), LG-Syntax User's Guide: Manual for Latent GOLD 4.5 Syntax Module. Belmont, MA: Statistical Innovations Inc.

Vermunt, Jeroen K. (2003), "Multilevel latent class models," Sociological Methodology, 33, 213-39.