SYMPOSIUM I ANVENDT STATISTIK

2025

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

Støttet af SAS Institute Inc.

Økonomisk Institut og Klinisk Institut
Syddansk Universitet

Symposium i Anvendt Statistik

Forord

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

Symposiet har til formål at understøtte delingen af statistiske analyser.

Dette års indlæg spænder over mange forskellige fagområder og lægger derudover vægt på metoder og analyser. Som det er normalt ved faglige 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 dialog for at udbygge kommunikationen mellem personer, der arbejder med beslægtede metoder inden for forskellige fagområder.

Peter Linde, Organisationskomiteen

ISBN 978-87-989370-5-0

Trykt hos PRinfoTrekroner i 80 eksemplarer

Organisationskomiteen for Symposium i Anvendt Statistik 2025

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

Peter Linde
Statistikkonsulent
Granparken 187
2800 Lyngby
peter@brede.dk

Anders Milhøj
Økonomisk Institut
Københavns Universitet
Øster Farimagsgade 5 – B26
1353 København K
Anders.Milhoj@econ.ku.dk

Esben Høg
Institut for Matematiske Fag
Aalborg Universitet
Thomas Manns Vej 23
9220 Aalborg Ø
esben@math.aau.dk

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

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

Helle M. Sommer
Konsulent
Hjortholmsvej 26
2830 Virum
helle.m.sommer13@gmail.com

Sara Armandi
Laerdal
Njalsgade 19D
2300 København
sara.armandi@laerdal.com

Nils Karl Sørensen
Institut for Økonomi
Campusvej 55
5230 Odense M
nks@sam.sdu.d

Klaus Rostgaard
Kræftens Bekæmpelse
Strandboulevarden 49
2100 København Ø
klar@cancer.dk

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

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

Hans Bay
Ekstern lektor
Københavns Universitet
Øster Farimagsgade 5
1353 København K
hans.bay@econ.ku.dk

Jane Rusbjerg
Data og Analyse
Kriminalforsorgen
Strandgade 100
1401 København K
Jane.Rusbjerg@krfo.dk

Arne Henningsen
Fødevare- og Ressourceøkonomi
Københavns Universitet
Rolighedsvej 23
1958 Frederiksberg
arne@ifro.ku.dk

Indholdsfortegnelse

Økonomi

Uddannelse

Sundhed II

Identifikation af blinde i danske sundhedsregistre

Katrine Prisak Jakobsen[1,2], Lonny Stokholm[1,2], Linda Juel Ahrenfeldt[3], Jakob Grauslund[2,4,5] og Sören Möller[1,2]

[1] Open Patient data Explorative Network, Odense Universitetshospital
[2] Klinisk Institut, Syddansk Universitet
[3] Forskningsenheden for Almen Praksis, Syddansk Universitet
[4] Øjenafdelingen, Odense Universitetshospital
[5] Steno Diabetes Center Odense, Odense Universitetshospital

Introduktion

Der formodes at cirka 20,000 borgere i Danmark er blinde eller stærkt svagtsynede, men det præcise antal er ukendt [2]. Blindhed er ifølge internationale studier associeret med ulighed i sundhed og livskvalitet, dette er dog ikke nærmere undersøgt i Danmark [4,1]. Vi planlægger derfor i projektet *AVID - Addressing Health and Socio-economic Disparities among Visually Impaired Individuals in Denmark* at undersøge ulighed mellem blinde og den resterende befolkning i Danmark på basis af nationale registerdata. Som første skridt i dette projekt ønsker vi at bestemme kriterier, der kan anvendes til at identificere blinde i danske registre, samt at estimere det samlede antal blinde i den danske befolkning.

Studiedesign og datakilder

Vores studie baserer sig på to datakilder, på den ene side et udtræk fra Landspatientregisteret (LPR) og CPR-registeret, som inkluderer alle 6,622,068 personer, der havde bopæl i Danmark og var i live 31.12.2022. På den anden side et udtræk af Dansk Blindesamfunds (DBS) medlemsdatabase med alle nulevende medlemmer på udtræksdatoen primo 2023, som hver især er bekræftet blind eller stærkt svagseende af en øjenlæge, som krav til indmeldelsen. På baggrund af disse to datakilder, har vi bestemt hyppigheden af øjnerelaterede ICD-10 diagnoser (samt tilsvarende ICD-8-koder) fra LPR både blandt DBS medlemmer, og bland den resterende befolkning.

Epidemiologiske resultater

I tabellen rapporterer vi den samlede gruppe af øjenrelaterede diagnostiske ICD-10-koder (DH*), samt de enkelte koder, der optrådte hyppigst i LPR for de personer, der indgår i DBS-datasættet. Vi observerer, at kun $6,4\%$ af DBS-medlemmerne er registreret med en kode, der indikerer dobbeltsidet blindhed (DH540), og $2,4\%$ med en kode, der indikerer stærk synsnedsættelse. Altså ville disse koder kun være i stand til at identificere mindre end 10% af DBS-medlemmerne som blinde, baseret udelukkende på LPR-data. På den anden side er $90,8\%$ af DBS-medlemmerne registreret i LPR med mindst en ICD-10-kode relateret til en øjensygdom. Altså har langt størstedelen af de blinde været i behandling i det danske hospitalsvæsen i forbindelse med øjensygdomme, for de fleste af dem dog uden

at være blevet registreret med en kode for blind/stærkt svagtseende i registrene. Derimod er der kun mindre end $0,1\%$ af ikke-DBS-medlemmerne, der er registreret med en kode, der indikerer blindhed / stærk svagtsynethed, som dog stadig giver 1344 individer, som må formodes at være blinde, uden at være medlem af DBS, og som derfor ikke kan identificeres via DBS-data alene.

	Dansk Blindesamfund	Resterende befolkning
Total	7.184	6.614.884
DH*	6.522 (90,8%)	960.090 (14,5%)
DH259	2.690 (37,4%)	238.699 (3,6%)
DH264	1.112 (15,5%)	53.341 (0,8%)
DH353	2.596 (36,1%)	71.305 (1,1%)
DH353J	1.094 (15,2%)	24.267 (0,4%)
DH353L	521 (7,3%)	8.868 (0,1%)
DH355	903 (12,6%)	2.130 (0,0%)
DH401	689 (9,6%)	16.593 (0,3%)
DH405	444 (6,2%)	4.679 (0,1%)
DH409	419 (5,8%)	9.130 (0,1%)
DH540	457 (6,4%)	815 (0,0%)
DH541	176 (2,4%)	529 (0,0%)
DH579	937 (13,0%)	71.694 (1,1%)

Capture-recapture-metoden

På basis af ovenstående resultater, er det tydeligt, at både LPR og DBS, kun kan identificere en andel af de blinde personer i den danske befolkning, og at der må forventes et stort mørketal af blinde, som ikke identificeres via hverken den ene eller den anden datakilde. Mens dette mørketal næppe kan identificeres på individniveau, giver capture-recapture-metoden (oprindeligt foreslået af den danske biolog C.G.J. Petersen, [3]) mulighed for at ekstrapolere størrelsen af dette mørketal, og dermed det samlede antal blinde i Danmark.

Vi vil i det følgende antage at sensitiviteten af en blinderegistrering (B) i DBS henholdsvis LPR er givet ved

$$P(DBS|B) = p^B_{DBS}$$
$$P(LPR|B) = p^B_{LPR}.$$

Samtidigt vil vi for nemheds skyld antage at specificiteten i både DBS og LPR er 1, altså

$$P(DBS|\neg B) = p^{\neg B}_{DBS} = 0$$
$$P(LPR|\neg B) = p^{\neg B}_{LPR} = 0$$

Mens dette næppe er fuldstændigt korrekt, så forventer vi at specificiteten er tæt på 1, da DBS kræver en undersøgelse hos en øjenlæge og en LPR-registrering for blindhed også antages at være af høj validitet.

Der vi altså gælde

$$P(DBS) = p^B_{DBS} \cdot P(B) + p^{\neg B}_{DBS} \cdot P(\neg B) = p^B_{DBS} \cdot P(B)$$
$$P(LPR) = p^B_{LPR} \cdot P(B) + p^{\neg B}_{LPR} \cdot P(\neg B) = p^B_{LPR} \cdot P(B).$$

Desuden vil der nødvendigvis gælde formelle sammenhænge

$$P(DBS\&LPR|B) = \alpha^B p^B_{DBS} p^B_{LPR} > 0$$
$$P(DBS\&LPR|\neg B) = \alpha^{\neg B} p^{\neg B}_{DBS} p^{\neg B}_{LPR} = 0$$

såfremt både $p^B_{DBS} > 0$ og $p^B_{LPR} > 0$, hvilket må antages, hvis datakilderne skal have nogen værdi.

På bagrund af Bayes formel kan vi derfor bestemme:

$$\begin{aligned}
P(DBS\&LPR) &= P(DBS\&LPR|B) \cdot P(B) + P(DBS\&LPR|\neg B) \cdot P(\neg B) \\
&= \alpha^B p^B_{DBS} p^B_{LPR} \cdot P(B) + 0 \cdot P(\neg B) \\
&= \alpha^B p^B_{DBS} p^B_{LPR} \cdot P(B) \\
&= \alpha^B \frac{P(DBS)}{P(B)} \frac{P(LPR)}{P(B)} \cdot P(B) \\
&= \alpha^B \frac{P(DBS) \cdot P(LPR)}{P(B)}
\end{aligned}$$

og altså (såfremt $P(B) > 0$)

$$P(B) = \alpha^B \frac{P(DBS) \cdot P(LPR)}{P(DBS\&LPR)}.$$

I vores datasæt observerer vi

$$\begin{aligned}
P(DBS) &= \frac{7184}{6622068} = 0,1085\% \\
P(LPR) &= \frac{457 + 176 + 815 + 529}{6622068} = 0,0299\% \\
P(DBS\&LPR) &= \frac{457 + 176}{6622068} = 0,0096\%
\end{aligned}$$

altså

$$\begin{aligned}
P(B) &= \alpha^B \frac{P(DBS) \cdot P(LPR)}{P(DBS\&LPR)} \\
&= \alpha^B \frac{0,001085 \cdot 0,000299}{0,000096} \\
&= \alpha^B 0,003379 = \alpha^B \cdot 0,3379\%
\end{aligned}$$

Og altså vil det samlede antal blinde n_B i befolkningen være

$$n_B = 6622068 \cdot 0,003379 \alpha^B = 22376 \alpha^B.$$

Hvis $\alpha^B = 1$, svarende til DBS og LPR værende uafhængige i deres identifikation, ville det indebære 22.376 blinde i Danmark.

Vi kan observere at konstanten α^B har denne relation til korrelationen mellem $(DBS|B)$ og $(LPR|B)$:

$$Cor(DBS|B, LPR|B) = (\alpha^B - 1) \frac{p^B_{DBS} \cdot p^B_{LPR}}{(1 - p^B_{DBS}) \cdot (1 - p^B_{LPR})}$$

Eftersom vi har observeret individer, der både er i DBS og LPR må $\alpha^B > 0$, men det kunne principielt tænkes at de to datakilder er stærkt negativt associerede, og at den totale andel og antal blinde derfor, som minimum, kunne være

$$P(B) = P(DBS) + P(LPR) - P(DBS\&LPR) = 0,1288\%$$
$$n_B = 6622068 \cdot 0,001288 = 8529.$$

svarende til, at der kun er de blinde i befolkningen, som vi har indentificeret i vores to datakilder.

I det andet ekstremum, kunne det tænkes at de to datakilder er stærkt positivt associeret, med en (næsten) vilkårlig stor værdi af α^B, og derfor en vilkårlig høj andel blinde i befolkningen. Dette er selvfølgeligt meget usandsynligt, men det er svært at give en øvre grænse for graden af positiv association.

Konklusion

Det viser sig at hverken DBS eller LPR i sig selv bidrager med nok information, til at kunne identificere alle, eller blot langt størstedelen, af de blinde i Danmark. Tilsammen giver disse to datakilder dog en formodentligt valid gruppe af blinde, der kan undersøges i registerstudier, dog med det forbehold, at der sandynligvis findes et mørketal på mindst 50% af de blinde i befolkningen med denne tilgang, hvilket det er vigtigt at være opmærksom på i fortolkningen af registerstudier på blinde.

Finansiering og taksigelser

Projektet er finansieret af Danmarks Frie Forskningsråd *AVID - Addressing Health and Socio-economic Disparities among Visually Impaired Individuals in Denmark* (10.46540/3165-00105B) og Gundhild Jenny Andersens fond (OUHs overlægeråds forskningsfond) *Identifikation af blinde i danske sundhedsregistre*. Vi takker desuden Dansk Blindesamfund for samarbejdet og tilladelse til at anvende et udtræk fra deres medlemsdatabase.

Referencer

1. Blinde og stærkt svagsynedes levevilkår. *VIVE – Det Nationale Forsknings- og Analysecenter for Velfærd* (2017).
2. Bengtsson, S., Mateu, N. C., and Høst, A. Blinde og stærkt svagsynede: barrierer for samfundsdeltagelse. *SFI - Det nationale Forskningscenter for Velfærd* (2010).
3. Petersen, C. G. J. The yearly immigration of young plaice into the Limfjord from the German Sea. *Report of the Danish Biological Station Band 6* (1896), 5–84.
4. Ulldemolins, A. R., Lansingh, V. C., Valencia, L. G., Carter, M. J., and Eckert, K. A. Social inequalities in blindness and visual impairment: a review of social determinants. *Indian J. Ophthalmol. 60*, 5 (Sept. 2012), 368–375.

Modelling height and weight in The Danish National Child Health Register

Gry Juul Poulsen,
Center for Molecular Prediction of Inflammatory Bowel Disease – PREDICT,
Department of Clinical Medicine,
Aalborg University

Summary:

In recent years new health Registries have been made available for research, such as Danish National Child Health Registry and the Register of Laboratory Results for Research, which include continuous variables measured at irregular and potentially informative time points. These data can allow researchers to study trajectories over time, but also pose certain challenges. In this talk I will discuss some of the challenges we have encountered in a study of growth trajectories before diagnosis of pediatric inflammatory bowel disease where we used height- and weight measurements from the Danish National Child Health Registry.

Background:

While most of the Danish health registries contain data on dates with categorical labels, such a date of hospitalization with associated diagnosis codes, some of the newer registries include data on continuous variables measured at a range of different time points. These datasets can potentially be used to study trajectories, for example of development in biomarkers before onset of disease, but the data have some shared challenges. The data are longitudinal but measured at irregular and potentially informative time points, and though the overall datasets may be huge, at an individual level the data can be sparse with potentially only a few observations per individual which can make modelling individual trajectories difficult.

One such resource is the Danish National Child Health Registry (DNCHR) which was established in 2009 and collects data on weight and height from health nurses and general practitioners along with data on breastfeeding and passive smoking in the home (Andersen 2023). Unlike birth cohort studies or other research projects where longitudinal data on children's height and weight are collected according to a pre-planned follow-up schedule, the data in the DNCHR are a mix of measurements from routine health checks and additional measurements taken because of concerns of the child's health, growth or weight.

Health registry data can address research questions that often cannot be answered with clinical databases such as long-term trajectories or trajectories before onset of a specific disease and have the advantage of covering all cases in the population rather than cases from a single specialized clinic. One example is inflammatory bowel disease (IBD) where registry studies have shown that patients show differences in medication use and biomarkers compared to population controls many years before they get an IBD diagnosis which suggests that inflammatory processes begin years

before onset of gastrointestinal symptoms (Vestergaard 2023, Bonfils 2023). For pediatric IBD, it is well-known that many patients have low weight and height for their age at time of diagnosis (Ishige 2019), but it is unknown how long before IBD diagnosis growth began to deviate. Brusco De Freitas 2025 examined how long before pediatric IBD diagnosis it was possible to identify differences in the mean height, weight and body mass index (BMI) by comparing Z-scores for height, weight, and BMI in up to 10 years prior to IBD diagnosis to up to 3 years after IBD diagnosis. In the following, I will discuss some the issues related to standardizing the height, weight and BMI measurements to a reference population and the problem of informative observation times.

Data:

The Danish National Child Health Registry (DNCHR) includes data on height and weight measured in infancy by child health nurses, at age 1-5 years at annual health checkups at general practitioners and at age 5-17 by school health nurses (Andersen 2023). During school age, measurements are taken at least three times: once in the year of starting school, once in middle school and once during last part of school. For some children more measurements are taken, in particular if there are concerns about the child's health, growth or weight. Figure 1 shows the distribution of age at measurements.

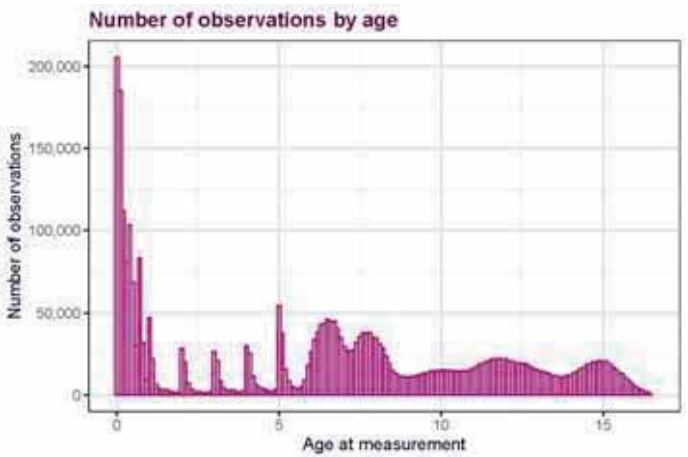


Figure 1: Distribution of number of height and weight observations in DNCHR by age at observation. Measurements are a mix of scheduled routine measurements and additional follow-up visits.

From the DNCHR, we identified a reference population consisting of 916,133 children born between 1997 and 2015 with at least one height/weight measurement during school age (between 5.5 and 16.5 years) and with birth length and weight recorded in the Danish Medical Birth Registry. We did not include observations in the first year of

life or measurements where either height or weight were outside +/- 5 standard deviation (SD) according to the 2014 Danish growth reference (Tinggaard 2014). After these exclusions the median number of measurements on each child was 3 with interquartile range (IQR) 4. Within this population, we identified an IBD population of 1,522 individuals diagnosed with IBD at age 5-17 years and available height and weight measurements between 10 years before and 3 years after first IBD contact. IBD was defined based on at least two inpatient hospital contacts/ outpatient visits with IBD (ICD10 code: K50, K51) within two years in the Danish National Patient Registry (Agrawel 2022). In addition to the IBD population, we identified a sibling population as a comparison group consisting of full siblings to the IBD population who did not themselves have IBD.

Methods

Standardization of data on height and weight

Since the aim of the study was to compare height, weight and BMI relative to time before and after IBD diagnosis, it was necessary to remove the dependence of age on height, weight and BMI. One way to do this is to calculate age- and sex-specific z-scores. If the underlying data are normally distributed, z-scores will have mean zero and standard deviation 1 and 99.9% of observations will be within -3 to 3. These properties make z-scores useful for comparisons across groups. However, if the data are not normally distributed, the z-scores will still have mean 0 and standard deviation 1, but the data are not necessarily within -3 and 3 and comparisons across groups can give misleading results if some groups have more skewed distributions than other groups.

The way that height, weight and BMI depend on age and sex is not straightforward: not only does the mean height and weight increase non-linearly with age, but for weight and BMI the standard deviation and skewness also changes non-linearly with age. One solution is to use an external growth reference such as the 2014 Danish growth reference (Tinggaard 2014). This growth reference was derived using methods that take age-dependent standard deviation and skewness into account, and the reference reports parameters in six-months interval that enable researchers to convert measurements into z-scores. However, because of the rapid growth during childhood using six-month means that children at the beginning of the age interval will get too low scores and at the end of the intervals too high scores, so using this reference would add noise to the data. Another solution is to use the DNCHR to create an internal reference; this is a sensible solution given the size of the DNCHR and has the further advantage that it eliminates the need of a non-IBD comparison group. Some authors have ignored the non-normality of data and simply standardized data by subtracting the mean and dividing by the standard error within intervals of age and by sex. However, as explained above, because the skewness in weight and BMI increases with age this approach might give misleading results as it does not completely remove the dependence of age. A better way to calculate z-scores is to use generalized additive

models for location, scale and shape (GAMLSS) to model height, weight and BMI as recommended by Borghi 2006. The GAMLSS model is a flexible, semi-parametric model that allows both mean, standard deviation, skewness and curtosis to depend non-linearly on covariates. We therefore fitted GAMLSS models with Box-Cox power exponential distributions for boys and girls separately including age as a continuous covariate using the lms function from the GAMLSS package in R (Stasinopoulos & Rigby 2008).

Selection of reference sample

The height and weight measurements in the DNCHR are a mix of scheduled routine measurements and additional measurements where either weight and height or other health problems gave rise to concern and additional follow-up visits. As example, children with at least one standardized BMI measurement greater than 2 according the 2014 Danish growth reference had a median of 5 (IQR 6) measurements compared with children who never had a standardized BMI measurement above 2 who had a median of 3 (IQR 3) measurements. Similarly, children with at least one standardized measurement below -2 had a median of 7 (IQR 6) measurements. This means that if all observations are used, children with either high or low weight for their age contribute with more observations than children with normal weight. For this reason, we chose to use one random measurement per child for the GAMLSS model to allow each child to contribute equally. This was done by drawing a random subsample consisting of 100,000 children using a weighted sampling upweighting measurements from age intervals with few measurements. The choice to use only 100,00 children out of the full reference population was made because of the long computation time of the lms and because the full reference population was much bigger than necessary. To examine whether taking a subsample of the reference population made any difference, we performed a sensitivity analysis where we drew ten different subsamples.

Analysis of growth in pediatric IBD patients and their siblings

Next the height, weight and BMI measurements of IBD patients and their siblings were standardized to z-scores which were modelled by years before and after IBD diagnosis using linear regression models including a random intercept for each child.

Results

Results from modelling growth by age and sex

Figure 2 shows the results of fitting the GAMLSS model with Box-Cox power exponential distributions to the height, weight and BMI data for boys and girls. In the figure actual measurements are coarsened to protect privacy and color is used to show the number of observations. For height the model fit the data well, except at 16 years where the curve unexpectedly begins to drop which may be an edge effect. For weight

and BMI, the fit is not quite as good: though the means are fairly stable the curves representing 2 and 3 standard deviations over the mean is less stable. Increasing the smoothing performed by the lms function did not help to make lines smoother. It is

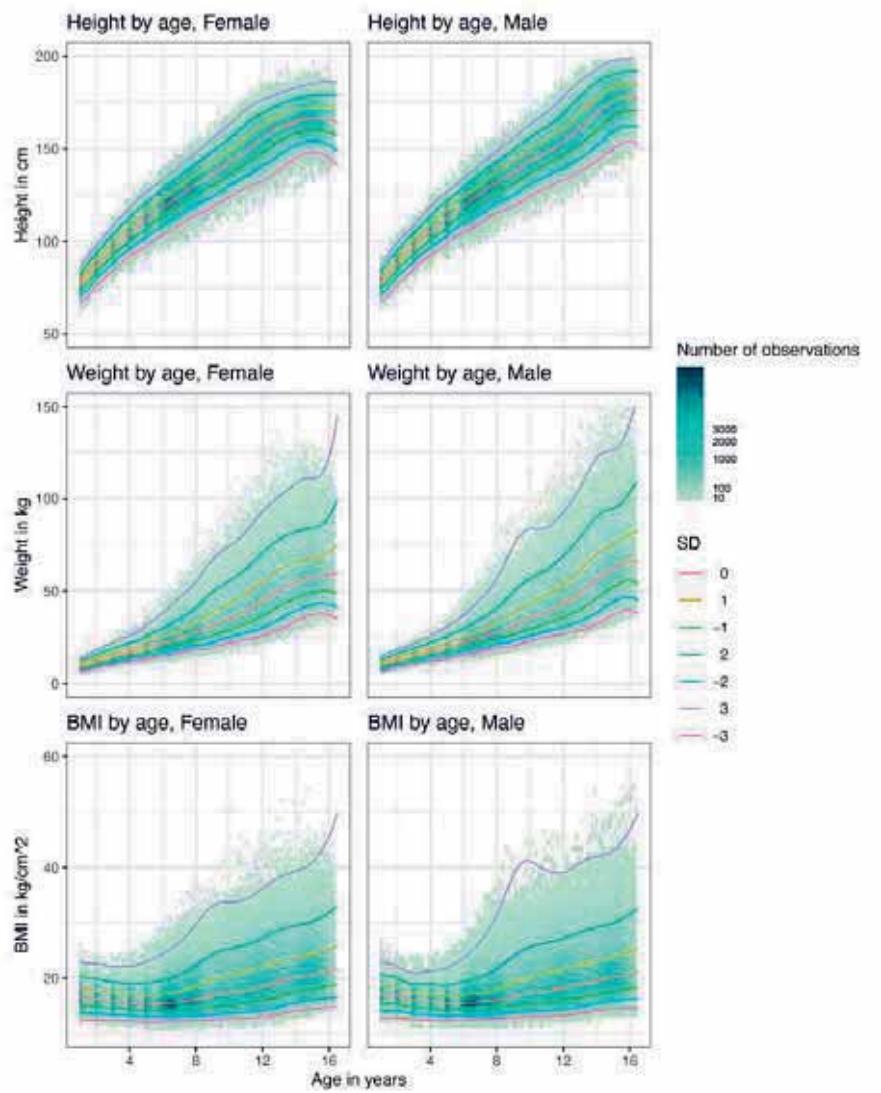


Figure 2: Fit of by GAMLSS models forheight, weight and BMI by age and sex. Underlying observations are coarsened to protect privacy and color is used to indicate number of observations in each cell. Lines represent means and 1, 2 and 3 standard deviations from the mean.

notable that the jumps in the 2 and 3 SD curves occur at age intervals where there are fewer observations; this could both be due to fewer observations in the age intervals or the fact that measurements in these intervals are less likely to be routine measurement and more likely to be extra measurements taken in children where there was a need to monitor height or weight more closely.

Figure 3 shows the three different methods applied on height and weight from children at age 5-16.5 and displays range of z-scores along with means and standard deviations in intervals of one tenth of a year. Using the 2014 Danish growth reference introduced small variations across six-moth age intervals for both height and weight. For height using the crude, untransformed z-scores and the z-scores derived from the GAMLSS model gave similar result, but for weight a substantial skewness was found in the crude, untransformed z-scores.

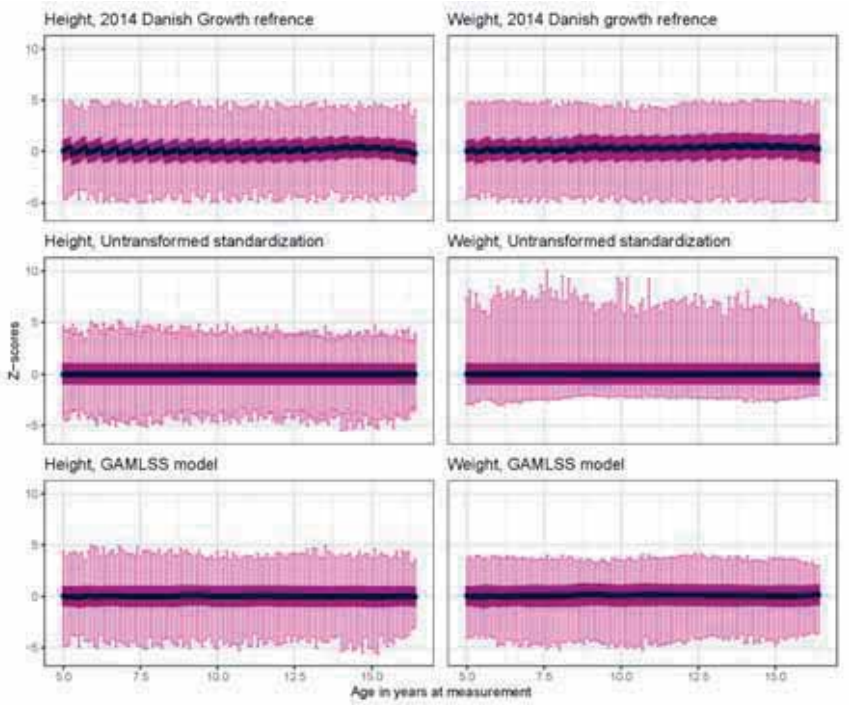


Figure 3: Results from standardizing height and weight on the reference population by three different methods: according to the 2014 Danish growth reference, by crude standardization without transformation and by GAMLSS models. The figure shows the range of the observations with mean marked by black dots and plus/minus standard deviation around the mean by dark purple. Age at measurement is divided into one tenth of a year.

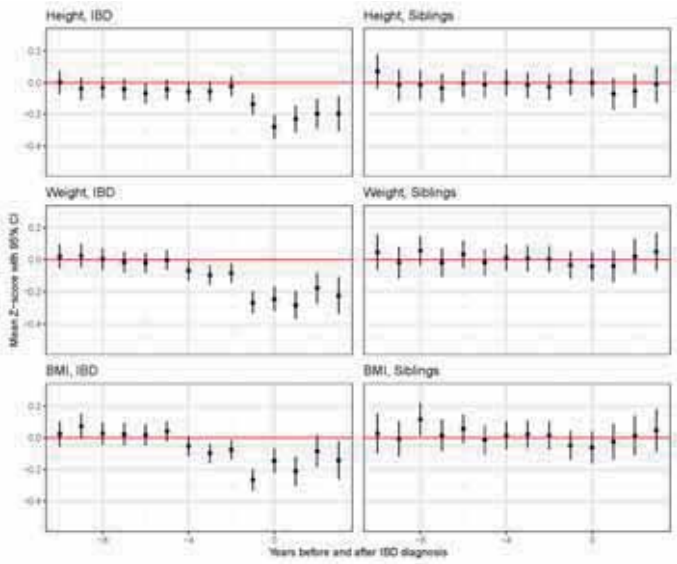


Figure 4: Mean z-scores with 95% confidence intervals for height, weight and BMI by years before and after IBD diagnosis for pediatric IBD patients and their full siblings. The sibling population is plotted by date of diagnosis of their IBD sibling.

Results from the analysis of growth in pediatric IBD patients and their siblings

The mean z-score for height, weight and BMI by time before and after IBD diagnosis is shown in Figure 4 and shows that children with IBD experienced a period of lower height, weight and BMI, both at year of diagnosis, but also in the years leading up the diagnosis and in the years after the diagnosis. For BMI the mean z-score was higher in the year of diagnosis than in the year before and after diagnosis. This may be due to systemic corticosteroids which 52% of the children received in year of diagnosis. By comparison, the siblings to IBD patients showed no departures from their normal height, weight and BMI through the period, which suggests that the changes in the BMI population is due to the onset of disease and not to family or environmental factors which would mostly be shared between full siblings.

Discussion:

Here I have presented some of the problems in the analyses of the height, weight and BMI measurements in the DNCHR. Correct standardization is important for the interpretation of results and for avoiding introducing noise or bias into the measurements while standardizing them. Further, using convenience data with a mix of routine and extra measurements introduce potential challenges if individuals with

high or low measurements have more observations. The height, weight and BMI measurements in the DNCHR are a good example of these challenges: the skewed distributions of weight and BMI need to be considered in the standardization, and because of height and weight is monitored at specific ages during childhood, it is to some extent possible to distinguish between routine and extra measurements.

As illustrated by the height in Figure 3, the crude standardization without transformation can be sufficient in some cases, and if the skewness does not depend on age it may also be sufficient to transform all measurements at once and then use crude standardization. For weight and BMI and other skewed measurements it is however necessary to use a more advanced method of standardization and the GAMLSS models are then a good choice. They are flexible and easy to use through the implementation of the lms function in the GAM package in R. However, there are some limitations. Firstly, the lms function is only able to include one continuous covariate; this means that the analyses did not take calendar year into account, though this will often be necessary in registry studies. Another limitation of using the GAMLSS models is the long computational time. While this is arguably in part a matter of sufficient computational resources and patience, it may also be a question of using data better and more efficiently. Cole 2021 provides a guide for designing studies of growth reference centiles and many of the considerations could also be applied when creating an internal reference.

In the construction of the internal reference for the standardization, we dealt with informative number of measurements by only using one observation per individual. Vogel 2017 discussed using resampling to deal with informative number of observations by using 1000 repeated subsamples to estimate reference curves. We tried to implement this approach in the current study, but again the size of the data was too big to make it feasibly to rerun the GAMLSS model so many times. However, we ran 10 resamples and found that it did not change results. While this approach reduces the impact of informative observation times, it is possible that problems with the fit for weight and BMI may be due to the age intervals without planned routine measurements. If children with observations in these intervals are likely to be children with either high or low weight, it might have been better to exclude these age intervals and base the model estimation on the age intervals with planned routine measurements. Informative number of measurements and measurement times may also play a role in the analysis of the IBD population. Since all children in the population are ill, it is likely that many of the children will have extra measurements which may be advantage for modelling trajectories. However, since more severely ill IBD patients may lose more weight and therefore have more measurements, the informative measuring times can still have an impact on results.

The discussion here has focused on height, weight and BMI in the DNCHR, but similar considerations would apply to studies of trajectories of biological measurements in the Register of Laboratory Results for Research. Many biomarkers

depend on age and sex, and using a GAMLSS model to create an internal reference can potentially be an advantage over using an external data (Vogel 2017).

References:

Agrawal M, Christensen HS, Bøgsted M, Colombel J-F, Jess T, Allin KH. The rising burden of inflammatory bowel disease in Denmark over two decades: a nationwide cohort study. Gastroenterology. 2022.

Andersen MP, Wiingreen R, Eroglu TE, Christensen HC, Polcwiartek LB, Blomberg SNF, et al. The Danish National Child Health Register. Clin Epidemiol. 2023;1087–94.

Bonfils L, Karachalia Sandri A, Poulsen GJ, Agrawal M, Ward DJ, Colombel JF, Jess T, Allin KH. Medication-Wide Study: Exploring Medication Use 10 Years Before a Diagnosis of Inflammatory Bowel Disease Am J Gastroenterol 2023;118:2220–2229.

Borghi E, de Onis M, Garza C, Van den Broeck J, Frongillo EA, Grummer-Strawn L, et al. Construction of the World Health Organization child growth standards: selection of methods for attained growth curves. Stat Med. 2006;25(2):247–65.

Brusco De Freitas M, Poulsen GJ, Jess T. Anthropometric trajectories in children prior to the development of inflammatory bowel disease. JAMA Network Open 2025 (in press).

Cole TJ. Sample size and sample composition for constructing growth reference centiles. Statistical Methods in Medical Research 2021, Vol. 30(2) 488–507.

Ishige T. Growth failure in pediatric onset inflammatory bowel disease: mechanisms, epidemiology, and management. Transl Pediatr 2019;8(1):16-22

Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. J Stat Softw. 2008;23:1–46.

Tinggaard J, Aksglaede L, Sørensen K, Mouritsen A, Wohlfahrt-Veje C, Hagen CP, et al. The 2014 Danish references from birth to 20 years for height, weight and body mass index. Acta Paediatrica, International Journal of Paediatrics. 2014;103(2).

Vestergaard MV, Allin KH, Poulsen GJ, Lee JC, Jess T. Characterizing the pre-clinical phase of inflammatory bowel disease. Cell Rep Med. 2023;4(11).

Vogel M, Kirsten T, Kratzsch J, Engel C, Kiess W. A combined approach to generate laboratory reference intervals using unbalanced longitudinal data. J Pediatr Endocrinol Metab 2017; 30(7): 767–773.

GDPR and other issues in reporting multiple outcomes to a common exposure

By Klaus Rostgaard

Danish Cancer Institute, Danish Cancer Society, Copenhagen

klar@cancer.dk

Abstract

It is a common and non-trivial task to devise a sensible categorization of a battery of disease outcomes to a common exposure. Obstacles/constraints include what can be communicated exactly (sufficient numbers of outcomes), what makes subject matter sense, what is relevant for the study, and changes in disease classification over the study, to mention a few. Here, we devise a simple semi-automated algorithm to deal with all mentioned obstacles.

Introduction – set-up and working example

The simplest and most common situation in which we report associations between multiple outcomes and a common exposure is when we communicate standardized incidence ratios (SIRs). A SIR $R_i=O_i/E_i$ is the ratio between observed (O_i) and expected (E_i) number of events (incidences of disease i) when following an exposed cohort, the exposure being e.g. a specific occupation or a history of a specific disease. The expectation E_i is usually calculated by summing products of sex-, age-, period- and cohort-specific incidence rates for disease i assessed in the national/regional cohort of people that the exposed cohort is nested within and correspondingly characterized time at risk in the exposed cohort [1]. The statistical analysis of SIRs is usually based on a Poisson model $O_i \sim Pois(R_iE_i)$, where the null hypothesis is $R_i=1$ and inference is about the parameter $\log(R_i)$ [1].

In practise the multiple outcomes we want to present typically represent a more or less complete non-overlapping categorization of the spectrum of diseases or cancer forms. Often we also want to present results for aggregated categorizations of outcomes, e.g. disease chapters (International classification of diseases – ICD) and organ systems (for cancers – ICDO) and finally for “any disease” and “any cancer”. Supplementary ad hoc groupings, e.g. smoking-related cancers, are also common.

As our working example throughout, we will use a purely register-based study by Andersen et al. [2] where we follow a cohort of 13,919 Danes with a hospital diagnosis of the skin disease hidradenitis suppurativa (the exposure), for the occurrence of a battery cancer forms (the outcomes), assessed in the Danish Cancer Register.

Challenges in presenting and analysing SIRs with low event counts

Data stewards, such as Statistics Denmark and the Danish Health Data Authority require that $O_i=0$ or $>x$, x typically 2 or 4, in order for them to be presented. When combining outcomes into a group of outcomes G we have $O_G \leq \sum_{i \in G} O_i$ and $E_G \leq \sum_{i \in G} E_i$ and de facto equalities when the numbers are small. Hence, we would very often have the data for small residual groups revealed if we present numbers for both a main group (an organ system) and all major cancers in that organ system. We also note, that there is nothing "informative" or surprising in observing $O_i=0$, unless E_i is substantially larger, which we could formalize in a criterion that E_i should be at least 3 or 5, symmetrical to the conditions for O_i. However, many applications of examining many outcomes at a time would be aiming only at circumstances of "too much", say in recipients of blood transfusions [3]. In such cases, the occurrences of $O_i=0$ are of limited interest. Furthermore, if O_i is just slightly less extreme we are not allowed to report it exactly. So, in the following, we will concentrate only on outcomes for which O_i is larger than some small integer x. For any(?) form of inference intended for an ensemble of outcomes, say a correction for mass-significance, it is a problem that the restrictions imposed by the data stewards are on the outcomes, the O_is, rather than what we know a priori, the E_is.

Another set of choices and challenges comes before reaching this point. Tradition has some classifications to offer, but changing classifications over a long time span of follow-up may not be entirely consistent and satisfactory. Furthermore, what is lumped together will often not be the same depending on whether the classification is geared towards the presention of diseases or the genetic causes of it [3–8].

For documentation and sanity, it should be a given constraint that any grouping employed should group together entities defined in other classifications.

One computationally good thing about this example (examining SIRs) is that you can figure out what groupings to use, only studying persons with events of potential interest. So, it can be made up of just tiny slivers of follow-up time around outcome events in the exposed cohort.

If the number of groupings under consideration are modest, as in our example of cancer forms, we can manually find out what groupings we prefer within the given constraints, as we did in [2]. If the goal is wider, say scanning essentially the complete spectrum of diseases [3], our approach preferably should be more algorithmic, for reasons of logistics and documentation.

A semi-automated bottom-up algorithm for generating presentable output

The above to us suggests a simple algoritm for generating the grouping of outcomes to be used that at all times stays compliant with revealing no O_i below some small number x, while at the same time leaving a few choices open to the investigator, when not all categories become large enough in the first two steps.

Algorithm

1) Generate O_i for all categories at the most detailed level, $O_i<x$ are represented as such.
2) Combine all categories within a main group for which $O_i<x$. Repeat 1) with these substitutions for small categories.
3) For combined categories for which $O_i<x$ decide to either 1) discard the category, including from the definition of the main group or 2) combine it with another category of your choice within the main group. Repeat step 2 with these new categories.
4) Define main group categories based on the categories in 3). A main group is dropped if the corresponding $O_i<x$.
5) Define the top level category ("Any cancer") as you please.
6) Calculate everything according to the union of the categories in steps 4) to 6).

References

1. Rostgaard K. Methods for stratification of person-time and events - a prerequisite for Poisson regression and SIR estimation. Epidemiol Perspect Innov **2008**; 5:16.
2. Andersen R, Rostgaard K, Pedersen O, Jemec GBE, Hjalgrim H. Increased cancer incidence among patients with hidradenitis suppurativa - a Danish nationwide register study 1977-2017. Acta Oncol **2024**; 63:220–228.
3. Dahlén T, Zhao J, Busch MP, Edgren G. Using routine health-care data to search for unknown transfusion-transmitted disease: a nationwide, agnostic retrospective cohort study. Lancet Digit Heal **2024**; 6:e105–e113.
4. Hebbring SJ. The challenges, advantages and future of phenome-wide association studies. Immunology **2014**; 141:157–165.
5. Wang L, Zhang X, Meng X, et al. Methodology in phenome-wide association studies: a systematic review. J Med Genet **2021**; 58:720–728.
6. Wu P, Gifford A, Meng X, et al. Mapping ICD-10 and ICD-10-CM Codes to phecodes: Workflow development and initial evaluation. JMIR Med Informatics **2019**; 7:1–13.
7. Pedersen MK, Eriksson R, Reguant R, et al. A unidirectional mapping of ICD-8 to ICD-10 codes, for harmonized longitudinal analysis of diseases. Eur J Epidemiol **2023**; 38:1043–1052.
8. Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annu Rev Biomed data Sci **2021**; 4:1–19.

Imputering af manglende blodprøver i laboratoriedatabasen

Frederik Lykke Petersen[1] & Sören Möller[1,2]

[1] Open Patient data Explorative Network, Odense Universitetshospital
[2] Klinisk Institut, Syddansk Universitet

1 Introduktion

Laboratoriedatabasens (lab-databasens) forskertabel består af mindst 1,9 milliarder prøvesvar fra landets klinisk biokemiske og klinisk immunologiske laboratorier, indsamlet siden 2008. Hovedparten består af blodprøver kodet som NPU-koder. Til berigelse af en population kan lab-databasen leveres til eksempelvis Danmarks Statistiks forskermaskine. Hvis forskeren ønsker anvendelse af tabellen i Stata på en stor population, kræves data-klargøring, primært pga. datastørrelsen (især hvis den komplet på hele befolkningen) og manglende standardisering af variable. I dette bidrag gennemgås nogle kodningsmæssige uklarheder i lab-databasen, og hvordan disse kan løses vha. SAS formater.

Som eksempel på analyse af mønstret i manglende blodprøver i lab-databasen på en population, tager vi udgangspunkt i det igangværende AI-HEMO projekt som er forankret på Aalborg Universitet i samarbejde med bl.a. Hæmatologisk afdeling, OUH og OPEN, OUH. I projektet analyseres sammenhængen mellem sjældne arvelige blodsygdomme og let tilgængelige blodprøver, som i fremtiden skal anvendes til at danne en prædiktionsmodel i lavindkomstlande. I en del af projektet defineres udfaldet som binært - havende en sjælden arvelig blodsygdom (N ca. 6000) eller ej, og disse matches med raske kontroller på alder og køn (N ca. 281.000), hvor indeksdato er casens dato for diagnosticering.

Prædiktorerne er otte forskellige prøver samt alder og køn. For hver relevant prøvetype har vi indsamlet den senest registrerede blodprøve før indeksdato og kun beholdt denne til analyse. I kohorten er symptomspecifikke blodprøvetyper mangelfulde på den symptomfri befolkning. Denne mangel er problematisk hvis variablene anvendes direkte som prædiktor for udfaldet. Til håndtering af mangelfuldheden kan anvendes imputering, eksempelvis i Stata.

En stærk antagelse i de følgende analyser er at mangelfuldheden er tilfældig (Eng: *Missing at random*). Hermed antages at mangelfuldheden er relateret til observerede variable, men ikke til uobserverede variable. I blodprøver er dette desværre meget usandsynligt, da prøvetagningsfrekvensen er højt korreleret til udfaldet, især for sjældnere symptomspecifikke blodprøvetyper. Med andre ord; I mange tilfælde er det kun patienter som er symptomatiske, som får specifikke prøvetyper taget, en form for *konfundering ved indikation* for mekanismen bag de manglende data, når udfaldet udelades af modellen.

2 Kodningsmæssige uklarheder – SAS formater

Data leveres i SAS format, og konverteres dette direkte til Stata, bliver alle værdier til tekstfelter som kræver megen diskplads. En optimeringsmulighed er at transformere

hver unik tekstbid en-til-en til et heltal og anvende SAS formater. For variable som har få unikke værdier er løsningen meget simpel. Eksempelvis findes variablen *operator* der angiver om den rapporterede måling er større eller mindre end de faktiske målte værdi.

```
proc format;
    value operator_label
    1 = "Storre end"
    2 = "Mindre end";
run;
```

De bagvedliggende heltal svarende til *operator* dannes vha. *if statements* og formateres med *operator_label* i et data-step.

```
data lab;
    set ekstern.lab_dm_forsker (keep operator);
    rename operator_new = operator;
    if operaotor = "stoerre_end" then operator_new = 1;
    if operator = "mindre_end" then operator_new = 2;
    drop operator;
    format operator_new operator_label.;
run;
```

Som supplement til prøverne leveres en oversigt over samtlige laboratoriekoder (*lab_dm_labidcodes*). Disse kan oversættes til et format i et data-step.

```
data lab_idcode_label (rename=(idcode=label));
    set ekstern.lab_dm_labidcodes (keep=idcode);
    start =_n_;
    fmtname = 'idcodefmt';
    type = 'N';
run;
```

For de resterende ikke-numeriske variable skal samtlige unikke værdier for variablene findes, før et formatbibliotek kan dannes. Eksempelvis *unit* som indeholder prøvernes måleenhed. Dette gøres i et SQL-step som derefter anvendes i et data-step der danner formatet som i ovenstående.

```
proc sql;
    create table unique_unit as
    select distinct unit
    from ekstern.lab_dm_forsker;
quit;

data unit_label;
    ...
run;
proc format cntlin unit_label;
run;
```

Data lab indeholder stadig *unit* tekstbidder, derfor skal de numeriske værdier kobles på i et SQL-step med *select*. Nedenfor en simplificeret udgave, kun på *unit*. *Bemærk at dette kan muligvis løses simplere.

```
proc sql;
    create table lab_merged
    as select A.*, B.start as unit
    from lab as A
    left join unit_label as B
    on A.unit = B.label;
quit;
```

I ovenstående skal * erstattes med samtlige variable der ønskes bibeholdt, bortset fra *unit*, eks. *A.var1*, *A.var2*. Husk at *B.start* er heltalsvariablen i formatet *unit_label*. Til sidst eksporteres data til Stata. Filerne opdeles på prøvekoder (DNK/NPU) vha. makroer. Koden hertil er udeladt.

3 Imputering af blodprøvedata

Vores blodprøvedata har ikke-monoton arbitrær mangelfuldhed. Dette vil sige, at variablene $X_1, X_2, \ldots, X_p$ ikke kan permuteres således at den næstkommende variabel i rækken har manglende værdier i samme observationer som foregående [1].

Flere muligheder til modellering af manglende værdier vha. imputation under ovenstående antagelse eksisterer, herunder prædiktiv middelværdimatchning (PMM) vha. K nærmeste naboer (KNN) og Bayesiansk multivariat normal regression (MVN). Paradigmemæssigt er disse to vidt forskellige. Et vigtigt mål for frekventister er ofte at kunne beskrive modeller eksakt vha. deres parameterisering. Dette behov kan i høj grad stilles ved at anvende MVN fremfor PMM da PMM modeller parameteriseres som en samling vektorer, dimensioneret på de uafhængige variable, og kræver relativt mere information end MVN, og modellens størrelse (i både parametre og pladsforbrug) derfor vokser med antal observationer og ikke blot med antal variable. MVN regressionsmodeller beskrives som på modelformen af en multivariat normalfordeling som

$$\mathbf{X} \sim MVN(\mu, \Sigma), \tag{1}$$

hvor X består af en observeret og uobserveret del. Dette tillader at alle uobserverede $X_i, i = 1, \ldots, n$ kan afhænge af observerede $X_i, i = 1, \ldots, n$.

Imputationerne dannes vha. en iterativ Markovkæde Monte Carlo (MCMC) model indtil konvergens er opnået. Detaljer herom er udeladt og kan læses i Stata dokumentationen [1]. I praksis udføres imputeringer i det følgende eksempel fra AI-HEMO projektet. For overskuelighed dannes modeller på samtlige variable nødvendigt, men kun variablene for målinger af hæmoglobin og RDW (Red Cell Distribution Width) analyseres i detaljer.

4 Imputering i Stata

4.1 Dataindhold

I Stata beskriver vi mangelfuldheden for målinger af hæmoglobin og RDW på individniveau vha Statas *misstable*.

```
. misstable sum hemoglobin rdw if case == 1
Variable     |   Obs=.      Obs>.      Obs<.
-------------+-------------------------------
hemoglobin   |   1,122                 4,768
        rdw  |   4,591                 1,299
-------------------------------------------------
. misstable sum hemoglobin rdw if case == 0
  Variable   |   Obs=.      Obs>.      Obs<.
-------------+-------------------------------
hemoglobin   |   167,932               113,353
        rdw  |   257,466               23,819
-------------------------------------------------
```

Første kolonne (Obs=.) viser at RDW er langt mere mangelfuld end hæmoglobin både for cases og kontroller.

4.2 Imputering

Vi udfører imputation af manglende variable vha. MVN. Variable som imputeres registreres som *imputed* og forklarende variable som *regular*:

```
. mi set wide
. mi register imputed hemoglobin rdw (...)
. mi register regular case age koen
```

Der imputeres, og modelparametrene for hver MCMC iteration gemmes i *ptrace.dta* indtil konvergens.

```
. mi impute mvn hemoglobin rdw (...) = i.koen age, ///
  saveptrace(ptrace)
```

For hver iteration i MCMC opdateres parametrene, og disse gemmes som en ny række i *ptrace*. Parametrene vil kunne anvendes til at trække stikprøver fra den konvergerede model, så ny data kan imputeres på den specificerede model. Stata har ikke en indbygget funktion til dette. Bemærk, at udfaldet er udeladt fra imputationerne, da det i praksis ikke vil være tilgængeligt, når der skal foretages prædiktioner på nye data. Det kunne være et interessant fremtidigt studie at sammenligne kvaliteten af imputationerne med og uden inkludering af udfaldet.

Vi vurderer nøjagtigheden af imputationerne ved at analysere hæmoglobin- og RDW-målinger.

```
. foreach v in hemoglobin rdw{
        misstable summarize `v', gen(mis_`v')
        mi xeq 1: summarize `v' if mis_`v' == 0; ///
                  summarize `v' if mis_`v' == 1; ///
                  summarize `v'
 }
```

For hæmoglobin:

```
m=1 data:
//For ikke-manglende data
-> summarize hemoglobin if mis_hemoglobin == 0
    Variable |   Obs        Mean    Std. dev.
-------------+--------------------------------------------------------
  hemoglobin |    156,632   8.57    1.08
//For manglende data
-> summarize hemoglobin if mis_hemoglobin == 1
    Variable |   Obs        Mean    Std. dev.
-------------+--------------------------------------------------------
//For manglende og ikke-manglende kombineret
  hemoglobin |   224,761    8.67    1.09
-> summarize hemoglobin
    Variable |   Obs        Mean    Std. dev.
-------------+--------------------------------------------------------
  hemoglobin |   381,393    8.63    1.09
```

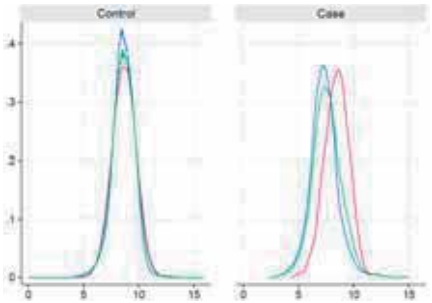
For RDW:

```
m=1 data:
//For ikke-manglende data
-> summarize rdw if mis_rdw == 0
    Variable |   Obs        Mean    Std. dev.
-------------+--------------------------------------------------------
         rdw |   33,248     .14     .40
//For manglende data
-> summarize rdw if mis_rdw == 1
    Variable |   Obs        Mean    Std. dev.
-------------+--------------------------------------------------------
         rdw |   348,145    .14     .40
//For manglende og ikke-manglende kombineret
-> summarize rdw
    Variable |   Obs        Mean    Std. dev.
-------------+--------------------------------------------------------
         rdw |   381,393    .14     .40
```

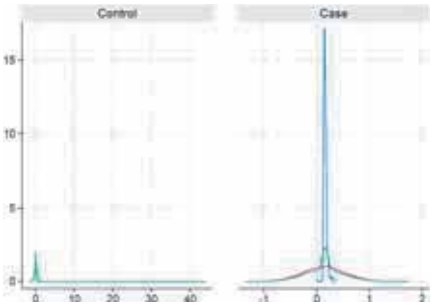
Ovenstående viser at middelværdier og standardafvigelser for imputationerne afviger ubetydeligt fra statistikkerne på observeret data.

Figur 1 og 2 illustrerer fordelingen af de observerede målinger, de imputerede målinger og de samlede målinger, stratificeret efter case/kontrol.

På figur 1 ses en god overensstemmelse mellem de faktiske og de imputerede værdier for hæmoglobin. Dette er dog ikke tilfældet for RDW (figur 2), hvor de observerede værdier for cases viser en meget smal fordeling, som imputationsmodellen ikke formår at gengive præcist. Denne snævre fordeling ville sandsynligvis blive bedre fanget, hvis udfaldet blev inkluderet i modellen.



Figur 1. Tætheder af hæmoglobinmålinger. Blå: Observeret, Rød: Imputeret, Grøn: Komplet



Figur 2. Tætheder af RDW-målinger. Blå: Observeret, Rød: Imputeret, Grøn: Komplet

Dette resultat er overraskende i lyset af *summarize*-tabellen, hvor middelværdier og standardafvigelser for de observerede og imputerede værdier var sammenlignelige. Det tyder på, at imputationerne formår at efterligne middelværdi og standardafvigelse, men ikke fanger højere momenter som skævhed og kurtosis, der kan være vigtige for at beskrive fordelingerne nøjagtigt. Derudover har imputationerne ikke en begrænset udfaldsmængde så imputationerne kan også være negative værdier, selvom disse ikke findes for RDW. Derfor bør valget af imputationsmodel for RDW genovervejes, så imputerede værdier bedre afspejler observerede værdier, hvilket muligvis kan gøres vha. Statas *mi impute truncreg* som tillader trunkering af udfaldsmængden. Det undersøges ikke nærmere om *truncreg* bedre kan beskrive fordelingen af RDW.

Litteratur

1. StataCorp. Stata multiple-imputation reference manual release 18. pages 11–12, 2023.

(Fejl)fortolkninger af konfidensintervaller

Tom Engsted

Institut for Økonomi, Aarhus Universitet

Fuglesangs alle 4, 8210 Aarhus V.

Email: tengsted@econ.au.dk

December 2024

Abstract: *Det er velkendt, at et klassisk 95% konfidensinterval, beregnet i en given stikprøve,* ***ikke*** *kan gives følgende fortolkning: "Med 95% sandsynlighed ligger den sande parameterværdi indenfor intervallet". Lærebøger anbefaler ofte alternative formuleringer á la: "Med 95% konfidens ligger den sande parameterværdi indenfor intervallet", eller "Med 95% sikkerhed ligger den sande parameterværdi indenfor intervallet". Hvad forstås egentlig ved 'konfidens' og 'sikkerhed' og hvorved adskiller de sig fra 'sandsynlighed'? Hvad skal man gøre, hvis man i en given stikprøve ønsker at beregne et rigtigt sandsynlighedsinterval for en parameter?*

Keywords: *Konfidensinterval; sandsynlighedsinterval; frekventistisk versus bayesiansk sandsynlighedsopfattelse; politisk meningsmåling.*

1 Indledning

Det klassiske dobbeltsidet konfidensinterval er et meget anvendt redskab til intervalestimation og – mere generelt – til angivelse af den statistiske usikkerhed omkring et punktestimat. Lærebøger pointerer som regel, at da en populationsparameter i klassisk statistik ikke er en stokastisk størrelse, kan et konfidensinterval ikke fortolkes som at det med en given sandsynlighed (typisk 95%) indeholder den ukendte populationsværdi. Det er netop derfor det kaldes for et 'konfidensinterval' og ikke et 'sandsynlighedsinterval'. Af og til anvendes også formuleringen at intervallet med så og så stor sikkerhed indeholder den sande værdi.

Lærebøgerne undlader dog som oftest at forklare, hvad der helt præcist skal forstås ved 'konfidens' og 'sikkerhed', og hvorved disse betegnelser adskiller sig fra 'sandsynlighed'. I en given stikprøve angiver et 95% konfidensinterval, at populationsparameteren med 95% konfidens (sikkerhed) ligger indenfor intervallet. Hvis konfidens (sikkerhed) ikke skal forstås som sandsynlighed, hvordan skal det da forstås? Og hvad skal man gøre, hvis man faktisk ønsker at beregne et rigtigt sandsynlighedsinterval for den ukendte parameter?

Disse spørgsmål diskuteres nedenfor, og der præsenteres en illustration med udgangspunkt i en politisk meningsmåling.

2 Det klassiske konfidensinterval

I det følgende tages der udgangspunkt i en normalfordelt population og en simpel tilfældig stikprøve, $y_1, ..., y_n$, hvor $y_i \sim N(\mu, \sigma^2)$, trukket fra denne population. Antag at vi er interesseret i populationens middelværdi, μ, og har estimeret stikprøvegennemsnittet, $\overline{y}$, og stikprøvevariansen, s^2. Et 95% konfidensinterval for μ fås da som

$$\overline{y} \pm \left(t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}} \right), \tag{1}$$

hvor $t_{0.975,n-1}$ er 97.5% fraktilen i t-fordelingen med $n-1$ frihedsgrader. Med tilstrækkelig stor n, kan t fraktilværdien approksimeres ved fraktilværdien i standard normalfordelingen.

For en given stikprøve, kan ovenstående konfidensinterval *ikke* fortolkes som at det med 95% sandsynlighed indeholder den sande værdi af μ. Årsagen er, at i klassisk (frekventistisk) statistik er populationsparametre ikke stokastiske størrelser, hvorved det naturligvis ikke giver mening at komme med sandsynlighedsmæssige udsagn om sådanne parametre. Stikprøveestimater af populationsparametre, derimod, er stokastiske, hvorved konfidensintervallets grænser bliver stokastiske. Ved gentagen stikprøveudtagning (*repeated sampling*) fås den statistiske fordeling for $\overline{y}$, hvor $s/\sqrt{n}$ angiver standardafvigelsen i $\overline{y}$'s fordeling.

Den korrekte fortolkning af konfidensintervallet bliver hermed, at hvis man gentagne gange trækker nye stikprøver fra populationen og beregner et 95% konfidensinterval i hver af disse, vil 95% af disse intervaller indeholde den sande værdi af μ. Man siger også, at intervallet har 95% *coverage probability*.

Det er vigtigt at understrege, at *for den givne stikprøve* fortæller konfidensintervallet intet sandsynlighedsmæssigt om parameteren. De 95% fortæller noget sandsynlighedsmæssigt om den *procedure*, der består i at beregne et 95% konfidensinterval gentagne gange i *forskellige stikprøver*. Jerzy Neyman ('opfinderen' af det klassiske konfidensinterval) var meget insisterende på den skelnen. I sin diskussion af, hvad "the practical statistician" kan udlede af et 99% konfidensinterval beregnet for en konkret stikprøve, siger han, at når denne "practical statistician" udleder, at populationsparameteren ligger indenfor intervalgrænserne, da gælder det at "in the long run he will be correct in about 99 percent of all cases." (Neyman, 1937, p. 349). I forlængelse heraf, siger Neyman: "the probability statements refer to the problem of estimation with which the statistician will be concerned *in the future* ... Consider now the case when a sample is already drawn and the calculations have given [the confidence interval]. Can we say that *in this particular case* the probability of the true value of [the parameter] falling between [the confidence limits] is equal to [x%]? The answer is obviously in the negative." (min kursivering). Et konfidensinterval beregnet i en konkret stikprøve enten indeholder eller indeholder ikke den sande værdi af parameteren. Ovenstående korrekte fortolkning af et konfidensinterval er naturligvis afledt af den særlige sandsynlighedsopfattelse, som kendetegner klassisk (frekventistisk) statistik. Neyman (1937) henviser flere steder til, at 'sandsynlighed' skal forstås som *long-run relative frequency* (i modsætning til den bayesianske sandsynlighedsopfattelse, se afsnit 4).

At det klassiske konfidensinterval alene siger noget sandsynlighedsmæssigt om en procedures egenskaber i *repeated sampling*, illustrerer dels de anvendelsesmæssige begrænsninger ved et sådant interval, samtidig med, at det åbner for diverse fejlfortolkninger i konkrete anvendelser.

3 Fejlfortolkninger

Som beskrevet ovenfor, giver et klassisk konfidensinterval ingen sandsynlighedsmæssig information om den ukendte parameter baseret på en konkret stikprøve. Lærebøger beskriver ofte et x% konfidensinterval – beregnet i en konkret stikprøve – på den måde, at det med x% *konfidens* (eller *sikkerhed*) indeholder den sande parameterværdi. Men lærebøgerne forklarer som regel ikke, hvad der helt præcist skal forstås ved 'konfidens' (eller 'sikkerhed'). Som vi har set, er disse betegnelser ikke blot alternativer til ordet 'sandsynlighed', men hvad er de så?

Når man – med de data, der udgør ens stikprøve – har beregnet et x% konfidensinterval for en parameter μ, ligger det snublende nær at fortolke "Med x% konfidens eller sikkerhed ligger den sande værdi af μ i intervallet" som substantielt identisk med "Med x% sandsynlighed ligger den sande værdi af μ i intervallet". Men brug af ordene 'konfidens' eller 'sikkerhed' i stedet for 'sandsynlighed' kan ikke skjule, at et klassisk konfidensinterval i en given stikprøve reelt intet inferensmæssigt fortæller om μ, men alene fortæller noget om egenskaberne ved *proceduren* med at beregne konfidensintervaller i gentagne stikprøver trukket fra populationen.

I en praktisk eller anvendelsorienteret kontekst, hvornår er en statistisk procedures sandsynlighedsmæssige egenskaber i gentagne anvendelser da interessante? Et eksempel herpå er statistisk kvalitetskontrol i produktionsvirksomheder. En fabrik producerer dimsedutter, der skal overholde diverse kvalitetsmål. Hver dag udtages en stikprøve af dimsedutter fra produktionen og kvaliteten testes. Her er den enkelte stikprøve ikke interessant i sig selv. Det interessante er i stedet, om man i den samlede produktion af dimsedutter på langt sigt sikrer sig, at andelen af fejlbehæftede dimsedutter holdes på præspecificeret niveau. De klassiske statistiske procedurer i form af konfidensintervaller og Neyman-Pearson hypotesetest er ideelle i sådanne sammenhænge. Til gengæld forekommer det problematisk at anvende disse procedurer, når data foreligger som passivt observerede ikke-eksperimentelle (*non-repeatable*) 'stikprøver', hvor formålet med analysen er at drage inferens om en parameter (eller model) baseret på ét konkret datasæt (Engsted and Schneider, 2024, diskuterer denne problemstilling mere detaljeret).

At det klassiske konfidensinterval alene siger noget om *coverage probability in repeated sampling*, indebærer samtidig, at man intet konfidens- eller sikkerhedsmæssigt (endsige sandsynlighedsmæssigt) kan sige om enkeltværdier eller delintervaller indenfor konfidensintervallet. Intuitivt vil vi ofte have den opfattelse, at værdier omkring midten af intervallet er mere 'sandsynlige' ('troværdige', 'plausible'; på engelsk: *likely*) end værdier tæt på intervalgrænserne. Men en sådan intuitiv opfattelse er fejlagtig. Som Cox (1958, p. 363) bemærker: "the method of confidence intervals ... gives only one interval at some preselected level of probability. ... when we write down the confidence interval ... for a completely unknown normal mean, there is certainly a sense in which the unknown mean θ is likely to lie near the centre of the interval, and rather unlikely to lie near the ends and in which, in this case, even if θ does lie outside the interval, it is probably not far outside. The usual theory of confidence intervals gives no direct expression of these facts."

Har man eksempelvis beregnet følgende 95% konfidensinterval for θ: [0.54; 1.12], og

nu gerne vil vide, med hvilken konfidens (sikkerhed) θ ligger i intervallet [0.90; 1.10], er man på Herrens mark; det klassiske konfidensinterval muliggør ikke besvarelse af dette spørgsmål (bayesianske sandsynlighedsintervaller, derimod, er direkte designet til besvarelse af sådanne spørgsmål, se afsnit 4).

Bemærk, at da intervalgrænserne er stokastiske (varierer fra stikprøve til stikprøve), er heller ikke følgende fortolkning korrekt (for ovenstående interval, [0.54; 1.12]): "Hvis vi gentager eksperimentet mange gange på nye stikprøver, da vil den sande værdi af θ i 95% af tilfældene ligge i intervallet [0.54; 1.12]".

The bottom line er, at et klassisk konfidensinterval, beregnet i en given stikprøve, ingen sandsynlighedsmæssig evidens indeholder om den underliggende populationsparameter. Om konfidensintervallet indeholder nogen form for 'evidens' overhovedet om parameteren, er et åbent spørgsmål! (jeg er tilbøjelig til at sige nej).[1] Hvis man er interesseret i sandsynlighedsmæssig evidens om en parameter – eller model – må man anvende bayesianske metoder, se næste afsnit.

4 Sandsynlighedsintervaller

Lad os nu i stedet opfatte populationsparametre som stokastiske størrelser, hvilket er hvad man gør indenfor bayesiansk statistik. Den sande værdi af en given parameter, μ, er ukendt, og vores á priori usikkerhed om værdien beskriver vi ved en sandsynligheds- eller tæthedsfordeling, $f(\mu)$. Informationen i data, D, opsummerer vi gennem likelihoodfunktionen $P(D \mid \mu)$, og á posteriori fordelingen for μ fås da – med anvendelse af Bayes' formel – som:

$$f(\mu \mid D) = \frac{P(D \mid \mu) f(\mu)}{\int P(D \mid \mu) f(\mu) d\mu}. \tag{2}$$

Á posteriori fordelingen $f(\mu \mid D)$ udtrykker vores usikkerhed om μ som en kombination af vores á priori usikkerhed og den information om μ som data giver os. Al statis-

[1] Af og til anvendes et konfidensinterval som mål for den præcision med hvilken en parameter estimeres. Men også dette er problematisk, se Morey et al. (2016). Generelt, indenfor frekventistisk statistik, skal statistisk usikkerhed opfattes som stikprøveusikkerhed, der måles ved spredningen i en statistiks stikprøvefordeling, eksempelvis $s/\sqrt{n}$, som er $\overline{y}$'s standardafvigelse, jf. afsnit 2. Jo mindre standardafvigelse, desto mere præcist bliver parameteren estimeret (dette naturligvis under antagelse af, at *repeated sampling* begrebet overhovedet giver mening, hvilket i høj grad kan diskuteres, når de underliggende data er passivt observerede ikke-eksperimentelle samfundsvidenskabelige data, jf. Engsted and Schneider, 2024). Når denne standardafvigelse dernæst anvendes til beregning af et x% konfidensinterval for μ med inddragelse af $\overline{y}$'s stikprøvefordeling, opstår alle de fortolkningsmæssige problemer beskrevet ovenfor.

tisk inferens om μ sker med udgangspunkt i $f(\mu \mid D)$, eksempelvis beregning af et x% sandsynlighedsinterval for parameteren (på engelsk kaldet *credible interval*).

Illustration: Politisk meningsmåling

Som illustration af forskellen mellem et klasisk konfidensinterval og et bayesiansk sandsynlighedsinterval, lad os se på den andel af befolkningen, der vil stemme på Enhedslisten (EL), hvis der var folketingsvalg. Antag, at seneste meningsmåling er målingen fra Megafon den 30. september 2024, ifølge hvilken EL står til at få 5.4% af stemmerne baseret på $n = 881$ respondenter. I målingen opgøres den statistiske usikkerhed for EL til $\pm 1.5\%$, der ifl. Megafon "angiver usikkerhedgrænsen for, hvor meget tallene i procentpoint kan svinge til hver side indenfor den statistiske usikkerhed ved et 95% konfidensinterval. Det vil sige, at den faktiske procentandel med meget stor sikkerhed ligger inden for det angivne interval".[2] (Bemærk anvendelsen af ordet 'sikkerhed' i stedet for 'sandsynlighed').

95% konfidensintervallet for EL beregnes konkret som (med antagelsen om en binomialfordelt stikprøve)

$$0.054 \pm \left(1.96 \cdot \sqrt{\frac{0.054 \cdot 0.946}{881}} \right) = [0.039; 0.069], \tag{3}$$

dvs. med 95% konfidens eller sikkerhed (ikke sandsynlighed!) ligger den sande andel, der vil stemme på EL, mellem 3.9% og 6.9%. Den helt præcise statistiske fortolkning er, at hvis man laver mange sådanne meningsmålinger, og beregner et 95% konfidensinterval i hver af dem, da vil 95% af intervallerne indeholde den sande andel, der vil stemme på EL; det *konkrete interval* i (3) baseret på én måling enten indeholder eller indeholder ikke den sande andel, og intervallet giver ingen sandsynlighedsmæssig information om andelen (andet end den trivielle, at sandsynligheden er enten 0 eller 1). I praksis udføres der løbende politiske målinger af forskellige meningsmålingsinstitutter, og politiske kommentatorer fortolker ofte den enkelte måling i sammenhæng med andre tilsvarende målinger. Dvs. *repeated sampling* begrebet er her ikke helt ved siden af, hvorved det klassiske konfidensinterval bliver et meningsfuldt redskab.

Men lad os nu beregne et 95% sandsynlighedsinterval for andelen (μ), der vil stemme på EL, og sammenligne det med konfidensintervallet i (3). Med en binomialfordelt

[2] https://politiken.dk/danmark/politik/art7158336/Se-den-seneste-Megafon-samt-%C3%A5rtiers-m%C3%A5linger-og-valgresultater?srsltid=AfmBOooxzB6bmYxRlWDE5fNbAnUx45d693xlHzWyVdht_Lx6yXksjRga

stikprøve, fås likelihoodfunktionen som $P(D \mid \mu) = \begin{pmatrix} n \\ X \end{pmatrix} \mu^X (1-\mu)^{n-X}$, hvor $n = 881$ og $X = 48$ (svarende til en stemmeandel på $48/881 = 0.054$). Lad os starte med at antage en uniform á priori fordeling for μ, dvs. $\mu \sim U(0,1)$, sådan at $f(\mu) = 1$. Det svarer til, at vi på forhånd ingenting ved om EL's sandsynlige stemmeandel. Indsættes i (2), fås hermed

$$f(\mu \mid D) = \frac{\mu^X (1-\mu)^{n-X}}{\int_0^1 (\mu^X (1-\mu)^{n-X}) d\mu} = \frac{(n+1)!}{X!(n-X)!} \mu^X (1-\mu)^{n-X},$$

hvilket svarer til en Beta($X+1$, $n-X+1$) fordeling. I det konkrete tilfælde bliver det en Beta(49, 834) fordeling. 95% sandsynlighedsintervallet bliver [0.0414; 0.0715], baseret på 2.5% og 97.5% fraktilværdierne i fordelingen.

Den uniforme á priori fordeling for μ kan forekomme urealistisk, da dens middelværdi er 0.50 (50%) og vi jo med ret stor sikkerhed ved, at EL vil få tættere ved 5% end 95% (eksempelvis) af stemmerne, hvis der var folketingsvalg. Lad os derfor antage en mere realistisk á priori fordeling. Ved sidste valg (i november 2022) fik EL 5.2%, og ved Megafons meningsmåling i august 2024 stod de til 7.6%. Dvs. udsving på plus/minus 2-3 procentpoint er ikke usædvanlig. Vælger vi derfor i stedet en Beta(6, 100) á priori fordeling, fås et mere realistisk billede af EL. Middelværdien i denne fordeling er 0.057, og 2.5% og 97.5% fraktilværdierne er hhv. 0.0213 og 0.1076. Á posteriori fordelingen $f(\mu \mid D)$ bliver en Beta(54, 933) fordeling[3], der har middelværdi 0.0547 og hvor 95% sandsynlighedsintervallet for μ bliver [0.0414; 0.0697].

Bemærk hvor numerisk ens de to sandsynlighedsintervaller er. Bemærk også hvor numerisk ens disse intervaller er med det klassiske konfidensinterval i (3). Det er et generalt resultat, at når stikprøvestørrelsen er relativ stor (som her, hvor $n = 881$), er sandsynlighedsintervallet ikke særlig følsom overfor det præcise valg af á priori fordeling, og i øvrigt numerisk tæt på det klassiske konfidensinterval. Forklaringen er, at når n er stor, dominerer informationen i data (likelihoodfunktionen) á priori fordelingen.[4] *Fortolkningen* af de to typer intervaller er dog væsensforskellig: Sandsynlighedsintervallet [0.0414; 0.0697] har den direkte fortolkning, at baseret på den givne stikprøve og den antagede á priori fordeling, er der 95% sandsynlighed for, at den sande andel, der vil

[3] Generelt, hvis á priori fordelingen er Beta(α, β), og stikprøven er binomialfordelt, bliver á posteriori fordelingen Beta($X+\alpha$, $n-X+\beta$).

[4] Dette står i skarp kontrast til hypotesetest, hvor resultatet af hhv. et klassisk test og et bayesiansk test af den samme nulhypotese afviger mere og mere fra hinanden, når stikprøvestørrelsen vokser (det såkaldte *Jeffreys-Lindley paradox*).

stemme på EL, ligger mellem 4.14% og 6.97%. Konfidensintervallet i (3) kan ikke gives en tilsvarende fortolkning; dette interval er blot ét ud af mange (hypotetiske) intervaller, om hvilke man ved, at 95% af dem indeholder den sande andel.

Á posteriori fordelingen $f(\mu \mid D)$ tillader endvidere beregninger såsom: *i*) Sandsynligheden for, at μ ligger mellem to vilkårlige værdier, eksempelvis $P(0.05 < \mu < 0.07)$. Baseret på Beta(54, 933) fordelingen kan denne sandsynlighed beregnes til 71.2%. Eller sandsynligheden for, at EL får max 7% af stemmerne, der kan beregnes til 97.7%. *ii*) De sandsynlighedsmæssige odds for, at EL får x% versus y% af stemmerne, $\frac{f(\text{x\% }|D)}{f(\text{y\% }|D)}$. Eksempelvis er odds for 5% versus 7% lig med $\frac{47.887}{6.497} = 7.37$, beregnet udfra Beta(54, 933) fordelingen, dvs. det er over 7 gange mere sandsynligt at EL får 5% af stemmerne, end at partiet får 7% af stemmerne. Den slags beregninger kan ikke laves med det traditionelle konfidensinterval.

5 Afsluttende kommentarer

Statistisk inferens baseret på det klassiske frekventistiske setup i form af konfidensintervaller og Neyman-Pearson (NP) hypotesetest, bygger på den sandsynlighedsopfattelse, at 'sandsynlighed' er en lang-sigts relativ frekvens i *(hypothetical) repeated sampling*. Type I og Type II fejlsandsynlighederne i et NP test er udtryk for lang-sigts frekvenser i gentaget anvendelse af testet, ligesom 95% *coverage probability* er det for et 95% konfidensinterval. I den forstand er de klassiske metoder *pre-data* procedurer, eller som Jerzy Neyman sagde: "the probability statements refer to problems ... with which the statistician will be concerned in the future".

I modsætning hertil er det bayesianske setup, hvor 'sandsynlighed' er udtryk for analytikerens *degree of belief* om en parameter eller model, herunder graden af usikkerhed om parameteren/modellen. Analysen er *post-data* i den forstand, at den giver sandsynlighedsmæssige udsagn om parameteren/modellen baseret på de faktisk observerede data.

Hvilken af de to tilgange der er mest hensigtsmæssig, afhænger af problemstillingen. Der er ingen tvivl om, at det frekventistiske setup er velegnet til eksempelvis statistisk kvalitetskontrol i produktionsvirksomheder, hvor antagelserne bag *repeated sampling* setup'et med rimelighed kan siges at være opfyldte, og hvor den enkelte stikprøve ikke i sig selv er interessant (og Neymans frekventistiske metoder blev da også udviklet til den slags problemstillinger).

Derimod er det svært at se anvendeligheden af disse metoder, når data foreligger som ikke-eksperimentelle og *non-repeatable*, hvilket er ofte forekommende indenfor samfundsvidenskab. Her kan de bayesianske metoder være mere appellerende, da de ikke bygger på en statistiks stikprøvefordeling i *repeated sampling*, men i stedet tager udgangspunkt i de faktisk observerede data, og hvad disse data inferensmæssigt kan fortælle om populationen.

I de fleste praktiske anvendelser er den slags overvejelser fraværende. De traditionelle metoder anvendes mere eller mindre ukritisk, uanset hvilken slags data man arbejder med. Når det drejer sig om ikke-eksperimentelle data, erklærer nærværende artikels forfatter sig enig med Kadane (2008, p. 457), der om de klassiske frekventistiske metoder siger: "They are based on the use of a procedure in a hypothetical infinite sequences of uses under the same circumstances. Thus classical inference relies on the idea that a single instance can be taken as typical of a hypothetical infinite population. I leave it to those who find this an attractive proposition to explain why."

6 Referencer

Cox, D.R. (1958): Some problems connected with statistical inference. *Annals of Mathematical Statistics* 29, 357-372.

Engsted, T., and J.W. Schneider (2024): Non-experimental data, hypothesis testing, and the likelihood principle: A social science perspective. *Foundations and Trends in Econometrics* 13, 1-66.

Kadane, J.B. (2008): Comment on article by Gelman. *Bayesian Analysis* 3, 455-458.

Morey, R.D., R. Hoekstra, J.N. Rouder, M.D. Lee, and E.J. Wagenmakers (2016): The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review* 23, 103-123.

Neyman, J. (1937): Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236, 333-380.

Measurement Error or Individual Variation

Associate Professor Emeritus, Department of Finance, Copenhagen Business School

Among other things, I will focus on some studies where preferences are measured at the individual level. Since we cannot look into the head of an individual, we must use indirect measurements. We therefore use a model. The model describes the assumed structure and enables a discussion of estimation methods, as well as a discussion of what is estimated and, not least, allows the model to discuss the interpretation of estimates. Whether estimates should be interpreted at the individual level or at the population level, and the difference will often reflect the purpose of the analysis.

I will give some examples of how it is important in relation to the interpretation of an analysis how estimates are interpreted.

Using higher dimensional space to find solutions of problems

Jacob Hjelmborg, The Faculty of Health Sciences, Department of Public Health, Epidemiology, Biostatistics and Biodemography, SDU

The classic idea of embedding data into some higher dimensional space to find solutions of problems, e.g., separation of features is key and has a modern aspect through extended kernel-based methods. We focus on a recent proposal of a data analysis framework that involves the theory of operator algebras. We will be reviewing embeddings of data into certain spaces that have lots of potentially useful structure, that is, embeddings into Hilbert C*-modules that have the reproducing kernel property. We intend to relate to aims of classical statistical notions as interaction, classification and dimension reduction with applications in health science.

Objektiv sensitivitetsanalyse for tidsvarierende konfundere

Andreas Kristian Pedersen[1,2], Anna Mejldal[2], Afsaneh M. Nejad[3], and Sören Möller[2,4]

[1]Klinisk forskningsafdeling, Sygehus Sønderjylland, Danmark
[2]OPEN, Odense Universitet Hospital
[3]Institut for Matematik og Datologi, Syddansk Universitet, Odense
[4]Klinisk Institut, Syddansk Universitet, Odense

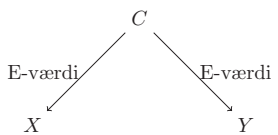
Resumé

I kausal inferens kan tidens rolle ikke overvurderes, da rækkefølgen af årsag og virkning er afgørende for at etablere kausale sammenhænge. I det kontrafaktiske udfald setup har tidsvarierende konfunderens og eksponering blevet undersøgt i stor udstrækning. Dog er det seneste gennembrud på dette område, E-værdien, ikke blevet udvidet til tidsvarierende konfunderens, udfald eller eksponeringer. Vi foreslår en udvidelse af E-værdien ved brug af den generelle opsætning af kontrafaktiske udfald, directed acyclic graphs, målteori og stokastiske integraler, hvor vi undervejs kun vil gøre minimale antagelser vedrørende fordelingen af konfunderen, i tråd med den oprindelige definition af E-værdien. Vi vil præsentere en stokastisk differentialligning for den umålte konfunder, der kan forklare den kausale sammenhæng, og vise, hvordan denne ligning kan løses numerisk under antagelsen om, at konfunderen er kontinuert over tid.

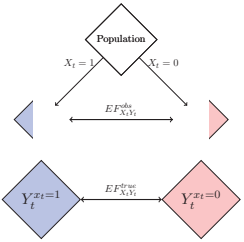
1 Introduktion

Konfundering er en af de mest problematiske størrelse indenfor biostatistik og epidemiologi hvor en variable C skaber en sammenhæng mellem to uafhængige variabler X og Y, der ofte er kaldt eksponeringen og udfaldet. Fænomenet ses ofte indenfor observationelle studier hvor man ikke kan fastslå om alt konfundering er taget højde for, og derfor ved man ikke hvorvidt den givne sammenhæng er reel. En metode er sensitivitetsanalysen E-værdien, der siger hvor stærk en given konfunder skal være associeret med eksponeringen og udfald på en risiko skala til at forklare denne sammenhæng man ser[DV16]. Problematikken er at E-værdien ikke er udvidet til longitudinelle målinger hverken i diskret eller i kontinuert tid, hvilket dette indlæg vil introducere.

For at kunne udvide E-værdien til kontinuert tid, skal den gennemsnitlige behandlingseffekt defineres. Vi bruger her Robins kontrafaktiske setup[HR06], hvor behandlingseffekten er defineret som forskellen mellem de to kontrafaktiske virkeligheder hvor alle var eksponeret eller ikke eksponeret, som beskrevet i figur 2.



Figur 1: E-værdi fortolkning



Figur 2: Definition of the effect sizes used in the derivation of the E-value process

Da eksponering ikke nødvendigvis er konstant over tid skal den gennemsnitlige behandlings effekt op til tidspunkt n defineres ved

$$\frac{1}{n}\sum_{k=1}^{n} Y_k^{X_k=1} - Y_k^{X_k=0},$$

hvor $Y_k^{X_k=i}$ er det kontrafaktiske udfald til tid k for $i = 0, 1$.

Derved ses der, under de klassiske antagelser indenfor kausal inferens, at vi kan estimere effektstørrelsen for den kausale og observerede sammenhæng i diskret tid på følgende måde[HR06]

$$EF^{true}_{X_tY_t} = \int \mathbb{E}[Y_t|X_t = 1, U_t = u]dF_{U_t\otimes U_{t-1}\otimes\cdots\otimes U_1}(u) - \int \mathbb{E}[Y_t|X_t = 0, U = u]dF_{U_t\otimes U_{t-1}\otimes\cdots\otimes U_1}(u)$$

$$EF^{obs}_{X_tY_t} = \int \mathbb{E}[Y_t|X_t = 1, U_t = u]dF^1_{U_t\otimes U_{t-1}\otimes\cdots\otimes U_1}(u) - \int \mathbb{E}[Y_t|X_t = 0, U_t = u]dF^0_{U_t\otimes U_{t-1}\otimes\cdots\otimes U_1}(u),$$

hvor U_t, X_t og Y_t er henholdsvis konfunderen, eksponeringen og udfaldet, $F_{U_t\otimes U_{t-1}\otimes\cdots\otimes U_1}(u)$ er det marginale mål af $U_t \otimes U_{t-1} \otimes \cdots \otimes U_1$, $F^i_{U_t\otimes U_{t-1}\otimes\cdots\otimes U_1}(u)$ er målet der korresponderer til den betingede fordeling af $U_t \otimes U_{t-1} \otimes \cdots \otimes U_1$ i gruppe $i = 0, 1$ og $EF^{true}_{X_tY_t}$ og $EF^{obs}_{X_tY_t}$ er henholdsvis effektstørrelsen for den ægte kausale sammenhæng og den observerede effektstørrelse.

Det ses nu at man kan definere både den ægte og observerede sammen mellem vores eksponerings og udfald ved at fokusere på funktioner af

$$f(t) = \sum_{i=1}^{n-1} \xi_i 1_{t_i,t_{i+1}}(t),$$

hvor ξ_i er $\mathcal{F}_{t_i}$-adapterede stokastiske variabler hvor $\mathcal{F}_{t_i}$ er filteret genereret af vores konfunder proces U_t til tid t_i. Det bemærkes at disse funktion approksimerer vores givne integrale i kontinuer tid, men ændrer sig i diskret tid. Derved får man følgende effektstørrelse i kontinuert tid når vores eksponering er binær

$$EF^{true}_{X_tY_t} = \int \mathbb{E}[Y_t|X_t = 1, U_{t,1}]dU_{t,1} - \int \mathbb{E}[Y_t|X_t = 0, U_{t,0}]dU_{t,0},$$

hvor $U_{t,i}$ er den betingede stokastiske proces af U_t for $X_t = i$ for $i = 0, 1$.

Det vil sige at hvis E-værdi processen defineres som forskellen mellem den kausale og observerede effektstørrelse har vi at den må opfylde følgende stokastiske differentialligning

$$dE(t) = EF_{U_{t,1}Y_t} f'(U_{t,0})dU_{t,0} + \frac{1}{4}EF_{U_{t,1}Y_t} f''(U_{t,0})d\langle U_{t,0}, U_{t,0}\rangle,$$

hvor $f'(U_{t,0})$ er associationen mellem konfunderen og eksponeringsvariablen, $EF_{U_{t,1}Y_t}$ er associationen mellem konfunderen og vores udfald, $f''(U_{t,0})$ er den afledte af associationen mellem konfunderen og eksponeringsvariablen og $\langle U_{t,0}, U_{t,0} \rangle$ er bracketfunktionen af $U_{t,0}$. Det bemærkes at både $f'(U_{t,0})$ og $EF_{U_{t,1}Y_t}$ har klinisk simple fortolkninger, der er styrken af konfunderen på henholdsvis en relativ og absolut skala for eksponeringen of udfaldet.

2 Løsning af differentialligningen

Når de givne størrelser er udfyldt i den stokastiske differentialligning ud fra den kliniske viden bemærkes det, at det kun er bracketprocessen af konfunderen, der sjældent er kendt når man arbejder med generelle semimartingaler.

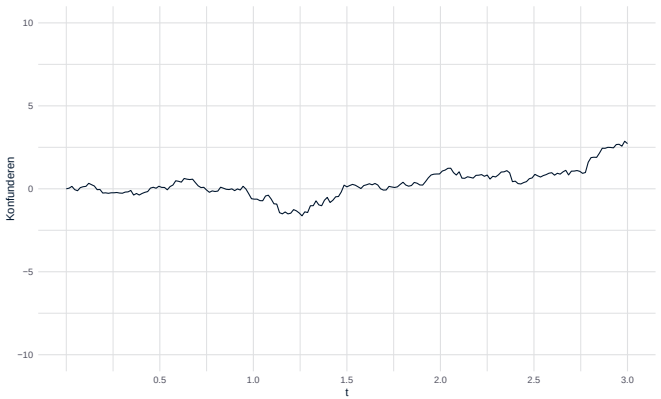
Det viser sig dog at hvis en konfunder ændrer sig kontinuert over tid kan denne approksimeres vha. Itô processer, hvorved bracketprocessen er givet ved

$$\sum_{i=1}^{n-1} \mu(t_i, U_{t_i,0}) d(t_{i+1} - t_i),$$

der giver følgende stokastiske differentialligning for konfunderen i kontinuert tid

$$dE(t) = EF_{U_{t,1}Y_t} f'(U_{t,0}) dU_{t,0} + \frac{1}{4} EF_{U_{t,1}Y_t} f''(U_{t,0}) \sum_{i=1}^{n-1} \mu(t_i, U_{t_i,0}) d(t_{i+1} - t_i).$$

Denne kan løses numerisk og derved kan den potentielle konfunder simuleres og følgende plot af konfunderen kan konstrueres



Figur 3: U_t over tid

Litteratur

[HR06] M. A. Hernan og J. M. Robins. "Estimating causal effects from epidemiological data". I: *J Epidemiol Community Health* 60.7 (jul. 2006), s. 578–586.

[DV16] P. Ding og T. J. VanderWeele. “Sensitivity Analysis Without Assumptions”. I: *Epidemiology* 27.3 (maj 2016), s. 368–377.

Violent aggression: Consequences of ostracism and violence against vulnerable adolescents

Mogens Nygaard Christoffersen, VIVE, The Danish Center for Social Science Research, mc@vive.dk

Abstract

Violent victimization and forms of ostracism of children and adolescents can have important developmental and psychological implications. Ostracism activates the same region of the brain as is activated by the experience of physical pain. Ostracism, meaning exclusion from the social context or group, can threaten fundamental needs, according to the need-threat model. An aggressive violence could be a response.

Method: Violent aggression by adolescents' and young adults is studied in a large-scale, prospective, longitudinal study of individuals born 1980-1985 followed in ages 15-27 years old (N=300,000) in linked records in Denmark. Violent aggression by young people, identified from police records as a charge of a violent crime is analyzed by a discrete-time log-odds model until an event occurs.

Violence against young people was identified in official records of parental violence (or domestic violence), child maltreatment, or victimisation of a violent crime.

The experience of ostracism is approximated from records of being adopted, being in care or being separated from family members. Not having graduated from high school, no vocational training, living in a disadvantaged area, or non-Danish citizenship are also taken as indicators of social exclusion.

Risk factors examined separately and jointly for their predictive power are: victimization of violence, and indicators of ostracism. Gender is controlled either as an intercept in a pooled model or by an interaction, splitting the sample into males and females

Results: Prevalence of aggression in young people was 4.8 % (n=14,607) as first-time charged up to age 27. Indicators of ostracism and violence in the family of the social environment preceded an increased risk of violent aggression among young people, both males and females.

It was not graduating from high school, or having no vocational training that most markedly elevate the risk of being charged of a violent crime.

Further, some vulnerable children known to be exposed to ostracism and violence (e.g. ADHD, brain injury or mental retardation) were found to be at higher risk of being charged of a violent crime.

There is also a huge variation between males' and females' charges of a violent crime for which we do not account. Generally, the model did not differ greatly between men and women, apart from Danish citizenship which appeared protective for men, but not for women.

Conclusion: Further study is needed of the mechanism where reaction to violence bears resemblance to that from pain inflicted by repeated and persistent experiences of ostracism from significant others in family, social groups, or local society, but also to what extent structural disadvantages and segregation will feed into higher levels of interpersonal violence. A need-threat model of ostracism provides a partial explanation of these associations.

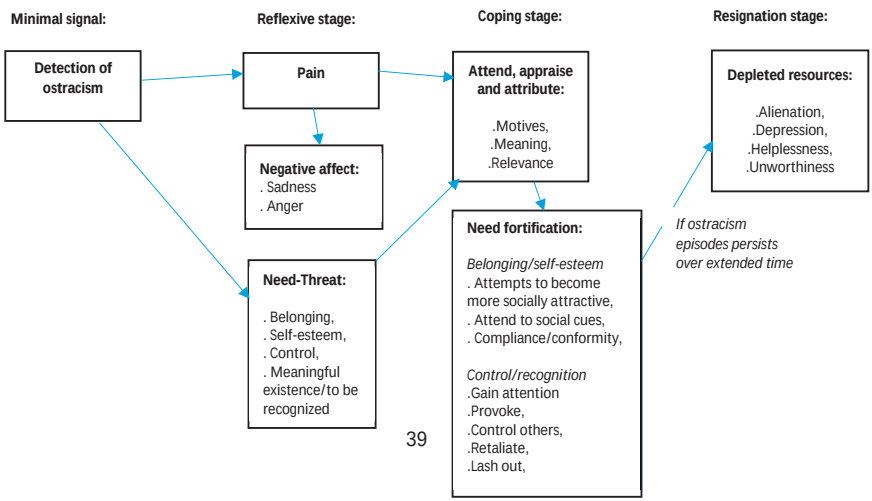
Introduction

Social exclusion comes in many shapes and forms. The political and economic structures and opportunities from which stratification and inequality emerge can produce social exclusion of the poor, undercut their pride and feeling of self-worth (Crutchfield & Wadsworth, 2003). Structural disadvantages reduce the opportunity of normal pathways to respect such as jobs, careers, and a capacity to be self-determining (Anderson, 2019). Structural features such as segregation can exacerbate the effects of poverty causing high levels of violent crime (Crutchfield & Wadsworth, 2003), and the effects of extreme disadvantage are pronounced for violent crime (Short, 2003). Both men and women living in a violent environment have a predisposition for violent aggression (Shaw & Dubois, 1995).

Williams' temporal need-threat model of ostracism

Why do we repeatedly find an association between previous violence in the family and local community and later violent behavior in adolescents?

Fig. 1. *Williams' temporal need-threat model of ostracism (Williams, 2009; 2022).*



Ostracism, meaning exclusion from the social context or group, can threaten four fundamental needs: the need to belong, the need to maintain self-esteem, the need to have control over one's social environment, and the need to be meaningfully recognized for existing. Williams' temporal need-threat model of ostracism (Williams, 2009) can serve as a theoretical framework for the study of long-term consequences of violence against children (Figure 1). The model proposes that reaction to ostracism happens in three stages: a reflexive stage; a coping stage; and a long-term resignation stage (Williams, 2009).

Ostracism is quickly detected and experienced as social painful and uncomfortable. The reflexive stage is seen as an adaptive response to a – for an evolutionary perspective – dangerous situation (Williams & Nida, 2022). Studies of ostracism in laboratory have shown heightened levels of activation in the same region of the brain that is activated by experience of the physical pain (Williams & Nida, 2022).

Not maintaining satisfactory levels of any of the four constructs (belonging, self-esteem, control over environment and recognition) may result in psychological harm in the coping stage. A child will engage in mental or behavioral activities to fortify the needs (Williams, 2009). According to Williams' model, the child will attempt to become more socially attractive; the child will try harder to adapt to relevant norms (Hutchison et al., 2007), or the child's learned strategy could be to gain control, provoke, retaliate or respond aggressively (lash out) in order to ensure adequate protection of the child's fundamental needs.

Under certain circumstances, retaliation, and generalized aggression can result from ostracism or social exclusion (Twenge et al., 2001; Warburton et al., 2006). Ostracism or social exclusion can be potential causes of aggressive behavior (Twenge et al., 2001). This is substantiated by the observation that many perpetrators of violence feel rejected by family members, peers, and society in general (Garbarino & Haslam, 2005; Leary et al., 2003; Twenge et al., 2001).

Can Williams' need-threat temporal model of ostracism also give us an explanation, if we find an increased incidence of violent aggression after young person is exposed to ostracism or violence?

Motivated by Williams' need-threat model we will narrow our study to recorded occurrence of ostracism or violence before first-time charges of violent crime, given other potential risk factors as uniquely measured for the population in national linked records. Can Williams' needs-threat temporal model of ostracism give us an explanation?

Social psychology in large scale studies

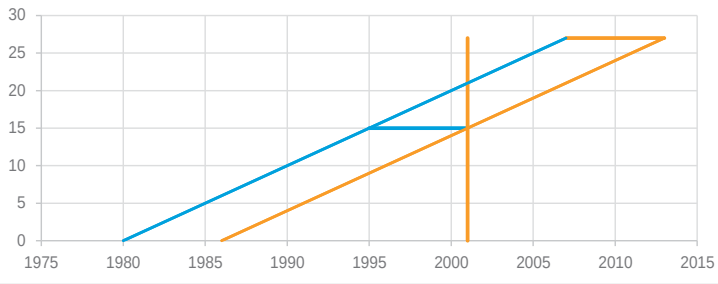
Method

The social psychologist focuses on individuals in face-to-face interactions (Tedeschi, 2003). The present study examines the implications for a child or young person of being exposed to ostracism and violence prior to being charged with a violent crime. The study is prospective and based on longitudinal data including the whole population sample, which provides us with information about the chronological sequence of potential causes before first-time registration of the events.

The total 6 birth cohorts (N=300,000) born between 1980 and 1985 are followed while they were 15 to 27 years old (Figure 2). Most data extending back to 1980 was available for both the cohort children and their parents. Hospitals records (diagnoses CD-8) were applicable from 1980, and (diagnoses ICD-10) were applicable from January 1994, while records concerning victims of sexual or violent assault were confined to years after January 2001, sourced from police records.

Prospective data about participants and their families was collected each calendar year from birth of the child. Risk factors represent developmental domains: inherited or acquired vulnerabilities, family stressors, exposure to family violence, parent-child relationship, indicators of ostracism, and disadvantaged living conditions, in general.

Figure 2. Lexis diagram. Birth-cohort of January 1980 to December 1985 being 15 to 27 years old in 2001-2012. Age by calendar year.

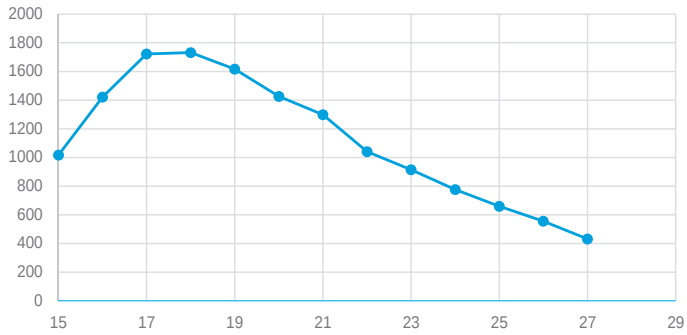


The events i.e. first-time charged of a violent crime against a person have been analyzed by log-odds models (Allison, 1982). The regressors were entered jointly and a procedure was carried out to select significant risk factors to give the best possible prediction. Individual event history is broken up into a set of discrete time units (a calendar year) in which an event either did or did not occur.

Analyzing the event history of an individual can be characterized as a sequence of events in order to answer the question: Why is it that some individuals get charged of a violent crime while others do not? The model we seek should incorporate explanatory variables that change in value over the observation period as well as variables that remains constant.

The discrete-time proportional odds model proposed by Cox (Cox, 1970) have several desirable features (Allison, 1982), but there are also problems in analyzing event histories. In the proportional hazard model (Allison, 2010), we assume that the ratio of the hazard for any two individuals is constant over time (age). Some age groups may be more developmentally vulnerable than other age groups. A vector $\boldsymbol{\alpha}_t$ is then added to the model in order to incorporate the age factor. We are not imposing any further restrictions on the set of constants (see Fig. 3). At age 18 years the highest numbers of first-time events are observed, but half of young persons have their first-time charges of violence after their 20th birthday.

Figure 3. Number individuals charged with a violent crime for the first time, by age. Individuals born between January 1980 and December 1985.



The national rules determining the age of criminal responsibility is 15 years in Denmark. Age 15 years is a natural starting point for our search. For each age group (age=15, 16, ..., 27) a constant (dummy) is estimated. It turns out that none of the time units has a frequency of 0. The smallest possible time unit is the calendar year for most of the time-dependent covariates. Each individual's event history is broken up into a set of discrete time units in which an event either did or did not occur. Then the data set consists of 3.8 million person-years.

All explanatory variables are treated as categorical dummy variables equal to one or zero. When the model was created (Allison, 1982; Cox, 1970) it was difficult and

time-consuming to analyze large data sets, but modern computers make it less costly to use the complete information. The proportional log-odds model allows in this way for standardization relating to age. Maximum likelihood estimators for the regression models are then calculated based on pooling of all the person years, and log-odds are presented (Allison, 1982). The log-odd discrete time methods allow us to estimate parameters in the model, treating each individual history as a set of independent observations. We can benefit from earlier findings where it has been shown that maximum likelihood estimators of parameters can be obtained by treating all the time units for all individuals as though they were independent (Allison, 1982).

The logistic regression function is written:

$$\text{Log}[P_{it}/(1-P_{it})] = \boldsymbol{\alpha}_t + \boldsymbol{\beta}'\mathbf{x}_{it}$$

P_{it} is the conditional probability that individual *i* has an event at age t, given that it has not already occurred to that individual. $\boldsymbol{\alpha}_t$ are estimated through dummy variables being 1 when the individual reaches the age *t*, otherwise 0. Both the dependent and the independent variables are dichotomized. Odds ratio is the ratio of the odds for x=1 to the odds for x=0. It follows that log odds ratio is $\boldsymbol{\beta}'$. Estimates of odds ratio will have a distribution that is skewed, in theory for large enough samples, the distribution will be normal. Log odds ratio follow a normal distribution for much smaller sample sizes (Hosmer & Lemeshow, 1989).

The model assumes that the ratios of the covariate's hazards remain constant (Allison, 2010). The explanatory variables $\mathbf{x}_{it}$ take on different value at different discrete age *t* (*t* = 15, … ,27). The observation continues until age t_i , at which point an event occurs or the observation is censored (Allison, 1982). Censoring means that the individual is not observed beyond the age t_i .

The model allows that individuals can have changing covariates over time, but the proportional hazards assumption means that the relative effect of each covariate is the same at all points in time (age). The coefficient we estimate is a sort of average effect over the range of ages in the data (Allison, 2010).

The time of observation of the explanatory variables may differ from actual time of the event in linked administrative data. Events are divided into three types of time-dependence. For example, the diagnoses of autism can be observed and diagnosed late in adolescence, but from theory about autism, we know that it has been present all the time. In Type I risk factors observed at time t also covers the years before and after t.

Some events will only influence the years when they are registered (Type II). In Type II we assume that the influence from being long-term unemployed will vanish when the redundant parent is back in employment. Type II describe a situation where the individual is exposed to a risk factor at time t that will only cover t, and perhaps t+1.

While other events (Type III) are assumed to cover all the following years, as for example a family separation.

Statistics

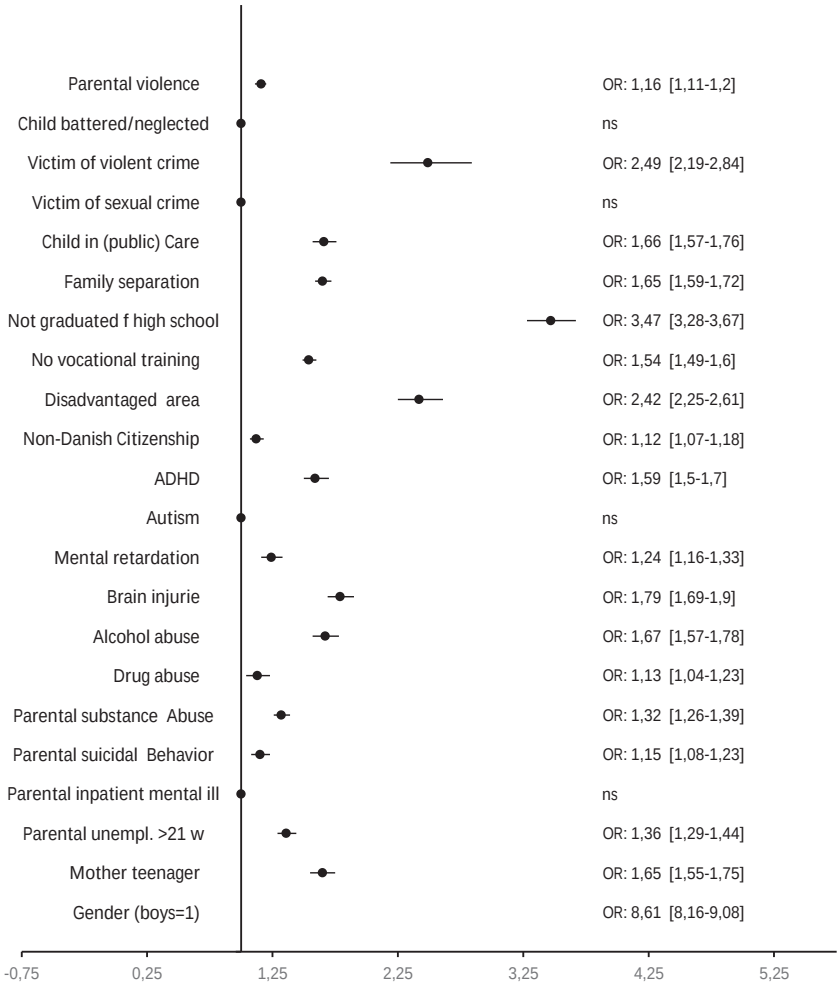
We wish to describe the environmental situation for the children the year before first-time event was observed in comparison with their contemporaries. The multivariate analyses included 23 risk variables (covariates) covering various aspects of the temporal need-threat model (Supplementary). Risk factors were included in the final model stepwise, if they contributed with new information, given all the others risk factors. The stepwise procedure was carried out to select the most significant risk factors to improve the prediction, and at the same time avoid redundancies. The stepwise selection is an automatic selection method (Peduzzi et al., 1980).

In order to evaluate the risk factors' contribution to the number of young persons registered with a charge of a violent crime, attributable fractions (AF) are calculated (Greenland, Sander, 1998). Attributable fractions express the reduction in incidence of the event that would be achieved if the population had not been exposed at all, compared with the current exposure pattern (Greenland, S. & Drescher, 1993). The estimated AF of a certain risk factor depends on two parameters only. One parameter is the strength of the risk factor measured by adjusted Odds Ratio (OR) or relative risk (RR). The other parameter is the current exposure of the risk factor in the population. The estimated AF is calculated solely on the basis on these two parameters (Levin, 1953; Woodward, 1999). Attributable fractions are only defined when OR and RR is more than 1. AF are not applicable to gender, because the population will always be exposed to both males and females.

The lifetime prevalence is estimated on the basis of the first-time incidence rate for the ages t=15, 16, 17, …, 27. We imagine that for instance 300,000 individuals live through the ages 15 to 27 years with the estimated age specific probabilities of first-time events.

The pooled model builds on an assumption that for example men and women range the seriousness of the experienced risk factors in the same order, at least. But what if for example men's response to risk factor A is an increased risk of violence, while women's response to the same risk factor is quite opposite. Analyzing men and women together would erroneously estimate the association between risk factor A and the following risk of being charged with a violent crime. We present separate analysis of males and females in order to test the assumption of proportional hazards.

Figure 4. Forest plot of risk factors. Charged of violent crime, adjusted Odds Ratio (OR), [95% CI]. Note: the plot of Boys OR: 8.61 [8.16-9.08] is not included in Fig. 4.



Men's and women's violent crime

In the data we observe that, according to police records 13,119 men were charged with a violent crime in comparison with only 1,488 women. The adjusted odds ratio for being male is 8.6 [CI 8.2-9.1]. We test how far the huge gap might be a consequence

of men and women experiencing different constraints and disadvantages and how far it could be a consequence of different responses to violence, ostracism and other disadvantages. From previous studies we know that more women react to violence and social exclusion with PTSD, suicide attempts and to a lesser extend with violence (Christoffersen, M. N. & Khan, 2024; Christoffersen, Mogens Nygaard & Thorup, 2024) .

Table 2.

Adjusted Odds Ratio estimates and Wald Confidence Intervals for men and women subsequently charged with a violent crime by age 27.

	Type	Men % of Controls	Men charged of violent crime OR	95% CI	Men AF %	Women % of Controls	Women charged of Violent crime OR	95% CI	Women AF %
Risk of violence									
Parental (domestic) violence	(I)	14.0	1.1***	[1.1-1.2]	1.4	15.8	1.6***	[1.4-1.8]	8.7
Child battered or neglected	(III)	0.5	Ns			0.6	Ns		
Victim of violent crime	(III)	0.5	2.4***	[2.1-2.7]	0.7	0.2	3.7***	[2.5-5.5]	0.5
Indicators of ostracism									
Child adopted	(I)	1.2	Ns	[2.9-3.3]		1.2	Ns		
Child in care	(III)	3.1	1.6***	[1.5-1.7]	1.8	2.8	2.1***	[1.8-2.4]	3.0
Family separation	(III)	33.5	1.6***	[1.6-1.7]	16.7	34.2	1.8***	[1.6-2.1]	21.5
Not graduated f high school	(I)	60.7	3.4***	[3.2-3.6]	59.3	50.1	3.6***	[3.0-4.3]	56.6
No vocational training	(I)	36.3	1.5***	[1.4-1.6]	15.4	28.9	2.0***	[1.8-2.2]	22.4
Disadvantaged area	(II)	1.7	2.5***	[2.3-2.7]	2.5	1.8	2.0***	[1.6-2.6]	1.8
Non-Danish Citizenship	(II)	7.8	1.2***	[1.1-1.3]	1.5	9.0	0.7***	[0.6-0.8]	NA
Selected disadvantages									
ADHD	(I)	5.2	1.6***	1.5-1.7	3.0	7.5	1.5***	[1.3-1.8]	3.6
Autism	(I)	4.7	Ns			7.0	Ns		
Mental retardation	(I)	3.6	1.2***	[1.2-1.3]	0.7	5.2	1.2*	[1.0-1.5]	1.0
Brain injury	(III)	4.6	1.9***	[1.7-2.0]	4.0	3.8	1.3*	[1.0-1.6]	1.1
Alcohol abuse	(III)	5.0	1.6***	[1.5-1.7]	2.9	6.7	2.2***	[1.8-2.6]	7.4
Drug abuse	(III)	2.2	Ns			3.6	1.3	[1.0-1.5]	1.1
Family risk factors									
Parental substance abuse	(I)	11.9	1.3***	[1.3-1.4]	3.4	12.4	1.2*	[1.1-1.4]	2.4
Parental suicidal behavior	(I)	4.0	1.1**	[1.0-1.2]	0.4	4.3	1.3***	[1.1-1.6]	1.3
Parental inpatient mental ill	(I)	11.4	Ns			11.8	Ns		
Parental unemployment>21 w	(II)	6.4	1.4***	[1.3-1.4]	2.5	6.4	1.4***	[1.2-1.7]	2.5
Mother teenager	(I)	2.9	1.7***	[1.6-1.8]	2.0	3.1	1.5***	[1.2-1.8]	1.5

Note: Numbers of men 13,119 numbers of women 1,488. Numbers of person-years in the study 1,898,626 and 1,932,175 for men and women, respectively. Ns means non-significant. AF means attributable fractions. NA means that AF is Not Applicable for odds ratio below 1.0. Type of dependency. Type I: Risk factor observed at time t also covers the years before and after t. Type II: Exposed to risk factor at time t are also present at t+1. Type III: Exposed to risk factor at time t, then the risk factor is also present at all the following years.

In most cases there are no significant differences between men's and women's exposure to known risk factors, or their response to them (Table 2). According to our measures in administrative records, young men and young women had been exposed to violence in the family to the same degree, but the women react more often

themselves with violent aggression, than women without such exposure. A relative high fraction of the charged women (AF: 8.7%) can statistically be linked to their exposure for parental violence (Table 2). The odds ratio was: 1.6; [1.4-1.8] for young women who had been exposed to parental violence to become violent themselves. Young men exposed to parental or domestic violence OR: 1.1; [1.1-1.2], also react with violence more often than other boys, but to a lesser extent than girls. The attributable fraction is only 1.4% among men.

We may conclude that estimating the strength (odds ratio) of being a member of a minority group (non-Danish citizenship) must be done separately for men and women, in order to get a non-biased result.

Strength and limitations

Prospective longitudinal studies following large population samples can help to avoid retrospective bias in self-reported survey data, but following individuals in birth cohorts have obvious methodological limitations. The temporal need-threat model raises the question of risk factors which are only partly or insufficient represented in the Danish Registry system.

We found that limited data exists on systematically registered violence against adolescents. The assumption is that the hazard rate is completely determined by the measured explanatory variables. In other words, we assume - naively you might say - that the vector $\mathbf{x}_{it}$ exhausts all the sources of individual variation in the hazard rate. The interaction between individuals and their environment differ in many ways which can possible capture all the variation among them (Allison, 2010).

Conclusion

Being a victim of violence and ostracism have important developmental and psychological implications in adolescence. Social rejection (ostracism), and violence against a child can increase the risk of later aggressive behavior as a response to get control over others. Williams' temporal need-threat model of ostracism provides a partial explanation of these associations. Reaction to violence bears resemblance to pain inflicted by repeated and persistent experiences of ostracism from significant others in family or peer group.

Victims exposed to violence or ostracism are at higher risk of future violent behavior. The number of adolescents charged of a violent crime were mainly associated with prior violence in the family, and family separation and structural ostracism. The included risk factors can only partly explain the occurrence of violent criminal behavior. The selected risk factors have a robust strength i.e. odds ratios significantly different from 1, but some of the risk factors are relatively rare in the Danish Registry system and therefore, they do not explain much of the aggressive violence. After all,

the robust associations between violence and the outcome of interest gives useful guidance for where to search for causal effects.

The results clearly indicate that victimization of violence and other forms of ostracism seems to have important developmental and psychological implications including violent aggression. The evidence that we can assemble for the need-threat model is clearly insufficient to explain violent aggression. The results also tell us that violence, ostracism, or loss of social support contribute to negative outcomes but only partly explains them. We would suggest that future research and other forms of data gathering also look for other risk factors and resilient factors in the environment.

FAIR data and Open Science: some insights from a work package in the PIGWEB project

Leslie Foldager

Dept. of Animal and Veterinary Sciences, Aarhus University, Tjele, Denmark
Bioinformatics Research Centre (BiRC), Aarhus University, Aarhus, Denmark
Email: leslie@anivet.au.dk

Abstract

There is an increasing demand and awareness of making (public) scientific research transparent and accessible both for other researchers but also to other levels of society, known as the Open Science movement. Pillars of Open Science include sharing of data, code, workflows, and publications but also transparency of peer reviewing. The PIGWEB project is a European project funded under the EU Horizon 2020 programme that aims to strengthen the pig research community by providing and facilitating transnational access to research infrastructures. One of the work packages focused on data sharing according to the FAIR principles: Findable, Accessible, Interoperable and Reuseable. This included getting an overview of facilities and awareness but also providing workshops on Open Science and FAIR principles, and on tools needed for FAIR data management and practices. Aspects of sharing via data papers and repositories were also presented and worth consideration.

Introduction

Sharing data and knowledge are elements of the Open Science movement as described in the UNESCO Recommendation on Open Science (UNESCO, 2021). Pillars of Open Science include sharing of data, code, workflows, and publications but also transparency of peer reviewing, and there is an increasing demand and awareness of making (public) scientific research transparent and accessible both for other researchers but also to other levels of society (Figure 1). Institutions are developing in-house strategies to promote data sharing and respect FAIR principles (Findable, Accessible, Interoperable and Reusable; Wilkinson et al., 2016), e.g., Data INRAE (https://data.inrae.fr/), RADAR in Germany (https://www.radar-service.eu/en) and Wageningen Data Competence Center (https://www.wur.nl/en/Value-Creation-Cooperation/WDCC.htm). Aarhus University also have instructions on the staff intranet as part of a policy endorsing the Danish Code of Conduct for Research Integrity (UFM, 2014).

The present proceedings paper presents some insights on FAIR and Open Science from work package 3 (WP3) of the EU Horizon 2020 project PIGWEB (An infrastructure for experimental research for sustainable pig production; https://www.pigweb.eu/) with partners from 10 European countries, including INRAE (National Research Institute for Agriculture, Food and Environment, France), WUR (Wageningen University & Research, The Netherlands), FBN (Research Institute of Farm Animal Biology, Germany), Agroscope (the Swiss confederation's centre of excellence for agricultural research), ILVO (Flanders Research Institute for Agriculture, Fisheries and Food, Belgium), SLU (Swedish University of Agricultural Sciences), IRTA (Institute of

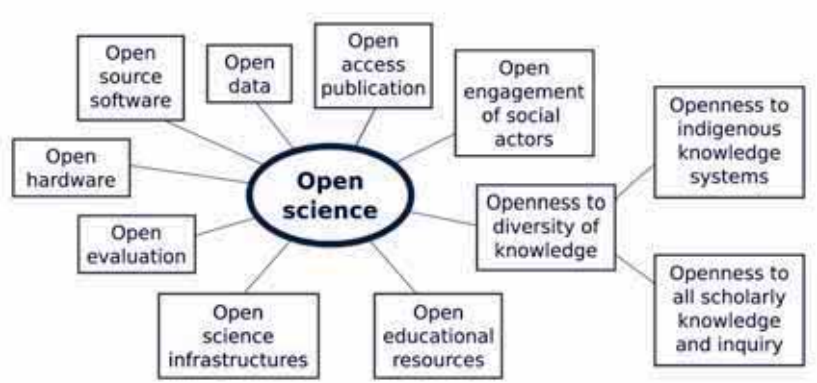


Figure 1. Open science elements.
Redrawn slide by Robbie Ian Morrison (20 Feb 2021, CC BY 4.0 via Wikimedia Commons) based on presentation by Ana Persic, Division of Science Policy and Capacity-Building (SC/PCB), UNESCO (France), Open Science Conference 2021, ZBW – Leibniz Information Centre for Economics, Germany.
https://commons.wikimedia.org/wiki/File:Osc2021-unesco-open-science-no-gray.png

Agrifood Research and Technology, Spain) and AU (Aarhus University, Denmark). The overall aim of the project is to strengthen the pig research community by providing and facilitating transnational access to research infrastructures, reinforcing a culture of cooperation between the research community and industrial and societal stakeholders, and improving and integrating the services provided by the research infrastructures. The project started in March 2021 and ends by February 2026.

In PIGWEB WP3, the focus is on FAIR sharing of data generated in the project. This includes getting an overview of facilities and awareness, but also providing training sessions on Open Science and FAIR principles, and on tools to support FAIR data management and practices. WP3 have provided two sets of training sessions on open science, data sharing and licensing, FAIR data management, metadata standards and annotation with ontologies (Table 1). The use of shared ontologies and metadata standards will improve reusability and machine-readability and allow faster and standardised data processing and analytics. Aspects of sharing via data papers and repositories are also worth consideration. Much of the contests in this proceedings paper and the subsequent talk is either inspired or (more or less) directly copied from the training session presentations – in a sense taking advantage of Open Science. The researcher mentioned in Table 1 are hereby credited with extra acknowledgement to the main organisers from WUR, Rob Lokers and Hendrik Boogaard.

Table 1. *Headlines from the PIGWEB WP3 training sessions on Open Science, data sharing and FAIR principles.*

Online seminar sessions (webinars), 29 Nov - 1 Dec 2022

Introduction to Open Science and FAIR principles (Rob Lokers, WUR)
Data management and the data life cycle (Hendrik Boogaard, WUR)
Data curation (Hendrik Boogaard, WUR)
Data publication (Rob Lokers, WUR)
Annotating data with Ontologies (Martin Toutant, INRAE)
Writing data management plans (Danny de Koning - van Nieuwamerongen, WUR)
PIGWEB data management plan (Catherine Larzul, INRAE)

On-site workshop, Wageningen, The Netherlands, 27-29 May 2024

Introduction and revisiting FAIR data practices and guidelines (Rob Lokers & Hendrik Boogaard, WUR)
Introduction to using ontologies (Catherine Hurtaud, INRAE)
- Animal Trait Ontology for Livestock (ATOL)[1]
- Environment Ontology for Livestock (EOL)[1]
- Animal Health Ontology for Livestock (AHOL)[1]

Progress on ABCD[2] metadata standard (Nina Melzer, FBN)
Understanding IPR[3] and data licensing - good practices and consequences (Rob Lokers, WUR)
Practical work on own data in small groups and discussions

[1] https://www.atol-ontology.com/en/atol-2/
[2] Access to Biological Collections Data
[3] Intellectual Property Rights

Open Science and FAIR principles

To cite the UNESCO Recommendation (UNESCO 2021), "Open science is a set of principles and practices that aim to make scientific research from all fields accessible to everyone for the benefit of scientists and society as a whole" and "Open science builds on the following key pillars: open scientific knowledge, open science infrastructures, science communication, open engagement of societal actors and open dialogue with other knowledge systems". This includes making publications, data, physical samples, software, and dissemination accessible to all levels, amateur or professional (wikipedia.org). Open science should encompass unhindered access to scientific articles, access to data from public research, and collaborative research enabled by ICT (Information and Communications Technology) tools and incentives (OECD, 2015).

The European Commission also actively supports the Open Science agenda and work according to the FAIR principles, and for example provides the *Interoperable Europe Portal* (https://interoperable-europe.ec.europa.eu/), to share knowledge on issues and

solutions. So, what is FAIR data sharing? **F**indable means that data should be discoverable with metadata, identifiable and locatable by means of a standard identification mechanism. **A**ccessible means that the data should always be available and obtainable; even if the data is restricted, metadata should be openly available. **I**nteroperable is perhaps the trickier part but means to be both syntactically parseable (i.e. possible to transform from one format to another) and semantically understandable so that data exchange and reuse is possible between researchers, institutions, organisations, or countries. **R**eusable means that data should be sufficiently described and shared with the least restrictive licences, thereby allowing the widest reuse possible and the least cumbersome integration with other data sources.

FAIR – why and how?

Reasons why we should comply with the FAIR principles include that we want to make use of the valuable work of peers, find and access their data and make sure they can find and access our data, and thereby avoiding reinventing wheels so that research is more efficient. By describing our data in a way that others understand and formatting it such that it fits the needs of other researchers and tools, we also want to increase the chance of getting credits for the work that we did to generate the data. Using FAIR data sharing also helps to safeguard our scientific integrity by transparency and verifiability, prevents loss of data, increases visibility, and drives innovation.

Sometimes we simply also must do it because funders require us to make the data reusable, institutions have it as part of policies, and journals encourage or even request data to go with the publication. As an example, the management at AU has adopted 3 strategic goals in relation the data sharing to be followed by all researchers at AU: 1) relate to the FAIR principles (data as well as to other outputs), 2) integrate data management into the research processes to ensure transparency and integrity in the results and 3) contribute to good practice and clear standards for handling data as well as metadata throughout the life cycle of the research. The last point includes data collection, curation and storing both during and after the completion of projects, including choosing licenses and using persistent indicators (PID), such as DOI (i.e. Digital Object Identifier; https://www.doi.org/).

According to a report from the EC Expert Group on FAIR Data (EC, 2018), for data to be FAIR, it needs to be represented in standard formats and accompanied by PIDs, metadata and code (EC, 2018). Identifiers should also be applied to other related concepts, e.g., identifying authors by ORCID (Open Researcher and Contributor ID; https://orcid.org/). Data should be represented in common and ideally open file formats and accompanied by the code used to process and analyse the data. It should also be accompanied by sufficient metadata and documentations to understand how, why, when and by whom the data were created, and have a clear accessible data usage license.

The FAIR principles may trigger some misunderstandings such as “I need to share all my data”, “Others might misuse my data”, and “I do not benefit from data sharing”. The use or reuse of the data can, however, be limited by specifying a suitable data licence and the data can be kept (partly) closed. The main requirement is to make sure that people know it exists and if or how they can get access. By providing good metadata the risk of misuse can be reduced. It can never be guaranteed that no one misuses the data, but the risk of at least unintended misuse can be reduced by describing the background and provenance of the data, and by providing code for data processing if needed and possible. Requiring attribution through a relevant licence increases the chance of being credited and thus benefit from the sharing of data. Moreover, publication of data papers in journals gives the opportunity to have the data sets published, peer reviewed and cited.

Data sharing

Sharing data is an important aspect of Open Science but is only useful if others can work with it; find the data, get the data, understand what it is about, and easily process it. There is however no requirement in Open Science for the data to be open. Data can also be shared with restrictions but should be as open as possible and only as closed as necessary. There may be reasons not to make data open; privacy (GDPR), contractual obligations, the volume (transfer/archiving costs), or simply irrelevant for reuse.

Data can be categorised into three types: 1) open data, 2) shared data and 3) closed data. Open data is data that anyone can access, use, and share. It must however be licensed to make clear that anyone can use the data in any way they want, including transforming, combining, and sharing it with others – and even for commercial purposes. Shared data may also be widely accessible but under some conditions such as non-commercial reuse. Not all shared data has to be available to anyone. At the other end, closed data could be highly sensitive data such as personal data or commercially sensitive data, where it may not be possible to share the data at all. Nevertheless, a metadata description of such research data should be shared. Often, a relevant Creative Commons (CC) license is applicable (https://creativecommons.org/share-your-work/cclicenses/).

Where to share or publish the data

When deciding where to share, deposit or publish the data it may be worth thinking ahead and consider perhaps some of the following. Maybe there is a repository hosting data from your research domain, which may then more likely also be used and searched by fellow researched from this field. Perhaps searching a registry of repositories can be helpful, such as re3data.org (https://www.re3data.org/). Check if the repository provides the data with a DOI or another PID and, if needed, whether restricted access to data can be provided. Some examples of repositories are *figshare* (https://figshare.com/), *DRYAD* (https://datadryad.org/) and the EC funded *Zenodo* (https://zenodo.org). Institutions may also offer repositories such as the AU facilities *REDCap* (https://redcap.au.dk/), a secure web platform for building and managing online databases and surveys, *ERDA*

(Electronic Research Data Archive; https://erda.au.dk/) for storing, sharing, analysing and archiving research data, and *SIF* (Sensitive Information Facility; https://sif.au.dk/) for storing sensitive data.

To figure where to deposit data or publish data papers, references to data sets in publications may also be consulted, journal websites and guides to authors, as well as your peers, data stewards or institutional guidelines. In the interest of open research, the Peer Community In (PCI) organisation (https://peercommunityin.org/) may also be an option. PCI is a non-profit organisation of researchers offering transparent peer review, citable recommendation, and publication of scientific articles in open access for free. Currently there are seventeen thematic PCIs such as PCI Animal Science, PCI Health & Movement Sciences, and PCI Mathematical & Computational Biology. Also many established journals have open versions (with publication fees) like *animal – open space*, which "fully embraces Open Science and its philosophy is that all reproducible research, the data linked to that research and the associated points of views of the authors will contribute to knowledge gain" (c.f. https://animal-journal.eu/animal-open-space/).

Funding

The PIGWEB project has received funding from European Union's Horizon 2020 research and innovation program under Grant Agreement No. 101004770. This publication reflects the views only of the author, and the European Union cannot be held responsible for any use which may be made of the information contained therein.

References

1. EC (2018). Turning FAIR into reality: final report and action plan from the European Commission expert group on FAIR data. European Commission: Directorate-General for Research and Innovation, Publications Office. https://data.europa.eu/doi/10.2777/1524.
2. OECD (2015). Making Open Science a Reality. OECD Science, Technology and Industry Policy Papers, No. 25, OECD Publishing, Paris. http://dx.doi.org/10.1787/5jrs2f963zs1-en
3. UFM (2014). Danish Code of Conduct for Research Integrity. The Danish Ministry of Higher Education and Science [UFM, Uddannelses- og Forskningsministeriet]. https://ufm.dk/en/publications/2014/files-2014-1/the-danish-code-of-conduct-for-research-integrity.pdf
4. UNESCO (2021). Understanding open science. UNESCO open science toolkit. https://doi.org/10.54677/UTCD9302
5. Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data **3**: 160018. https://doi.org/10.1038/sdata.2016.18

Hvor liberale er Liberal Alliances vælgere?

Anders Milhøj

Anders.milhoj@econ.ku.dk

I efteråret 2023 blev der gennemført en lokal dansk udgave af European Social Survey (ESS), som er en indsamling udført med stikprøver af befolkningerne i de fleste europæiske lande. Desværre indgår Danmark ikke i disse år i de fælleseuropæiske dataindsamlinger, der foretages ca. hvert andet år.

Den lokale indsamling er finansieret af Økonomisk Institut, Københavns Universitet, men den følger i det store hele den fælles spørgeramme. Undersøgelsen, der blev gennemført af analysebureauet Wilke, havde ca. 1500 respondenter på 18 år og ældre, dvs. at den er repræsentativ, hvad angår køn, alder og geografi, som andre danske meningsmålinger.

Der spørges i den fælles ESS bl.a. om respondenten stemte ved sidste folketingsvalg samt hvilket parti, respondenten i givet fald stemte på. Desuden spørges om, hvilket parti respondenten føler sig tættest på. I den lokale stikprøve blev der yderligere stillet enkelte aktuelle spørgsmål, bl.a. om alkohol- og tobaksvaner samt om holdninger til aldersgrænserne for unges køb af alkohol- og tobaksprodukter.

Tilslutningen til Liberal Alliance

I stikprøven havde Liberal Alliance en tilslutning på 7.6% blandt de respondenter, der stemte ved valget i november 2022, hvilket er stort set det samme som resultatet 7.9% ved valget. I efteråret 2023, altså omkring et år efter valget, var tilslutningen lidt højere, 8.2%, hvilket er noget lavere end set i mange andre meningsmålinger fra efteråret 2023.

Blandt Liberal Alliances vælgere er der i denne undersøgelse en overvægt af unge, mænd, byboere og personer med højere indkomster, hvilket mange andre undersøgelser også har vist, se fx. Altinget Hvem stemmer på Liberal Alliance? Her er partiets typiske vælger - Altinget - Alt om politik: altinget.dk.

Alkohol- og tobaksforbrug

Det er velkendt, at netop spørgsmål om drikke- og rygevaner kan være svære at opnå pålidelige svar på, da mange respondenter forsøger at gøre besvarelsen mere 'korrekt', end den er i virkeligheden. I modsætning til tidligere ESS-undersøgelser, der var indsamlet ved face-to-face interviews, er denne ESS indsamlet via internettet, så måske er trangen til at lyve sig bedre, end man er, ikke så stor.

Tabellen viser alkoholforbruget for Liberal Alliances vælgere sammenholdt med andre partiers vælgere. Her er naturligvis kun brugt de respondenter, der stemte ved sidste folketingsvalg, hvilket fx. 18-årige respondenter ikke havde mulighed for.

Drikker du alkohol?

	Stemte på Andet	*Stemte LA*	*Total*
Slet ikke	221	11	232
	16.57%	11.83%	
En gang om måneden	412	34	446
	30.88%	36.56%	
En gang om ugen	521	39	560
	39.06%	41.94%	
Dagligt	164	9	173
	12.29%	9.68%	
Ønsker ikke at oplyse	16	0	16
	1.20%	0.00%	
Total	1334	93	1427

Det er tydeligt, at LA vælgernes drikkevaner ligner vanerne hos de øvrige partiers vælgere, da ingen forskelle er væsentlige. De små forskelle kan let forklares ved at LA's vælgere i hovedtræk er yngre end andre partiers vælgere.

Holdning til aldersgrænser for unges køb af alkohol

I de senere år har diskussionerne om aldersgrænserne for unges køb af alkohol og nikotinprodukter spidset til, da flere og flere røster ønsker højere aldersgrænser og en styrkelse af kontrollen med de indførte grænser. Derfor blev der i efteråret 2023 talt en del om disse aldersgrænser, og der blev indgået politiske aftaler.

Midt i november 2023 blev der indgået en såkaldt 'forebyggelsesaftale' mellem regeringen, Alternativet, Danmarksdemokraterne, Konservative og SF. Aftalen indeholdt skarpere aldersgrænser for salg af alkohol til unge, som blev vedtaget i juni 2024. Fra 1. juli 2024 må unge under 18 år kun købe alkohol med en styrke under 6%, svarende til højst guldøl, men ikke stærkere øl, vin eller de meget populære shots med en procent under den tidligere grænse fra 2010 på 16.5% for unge over 16 år.

Aftalen omhandlede tillige tiltag for at mindske salget af nikotinprodukter målrettet yngre, fx. forbud mod visse tilsætningsstoffer til e-cigaretter; dette mere komplekse forslag er i høring frem til august 2024.

Tabellen viser holdningen til aldersgrænser for unges køb af alkohol blandt Liberal Alliances vælgere sammenholdt med andre partiers vælgere. Der spørges altså til de grænser, der gjaldt i efteråret 2023, dvs. før skærpelsen i sommeren 2024.

Er aldersgrænserne for unges køb af alkohol…

	Stemte på Andet	*Stemte på LA*	*Total*
For lempelige / løse	451	21	472
	33.81%	22.58%	
Passende	863	70	933
	64.69%	75.27%	
For strenge	20	2	22
	1.50%	2.15%	
Total	1334	93	1427

Det vil sige, at 22.58% af LA vælgerne synes, at reglerne er for lempelige, mens hele 33.81% af de respondenter, der stemte på andre partier, synes grænserne er for lempelige. Forskellen er synligt stor, men dog kun lige knapt signifikant, $p = 8.1\%$. Hvis de få, der synes, at reglerne er for strenge, slås sammen med den store gruppe, der svarer, at reglerne er passende, bliver $p = 2.2\%$, dvs. at forskellen er signifikant.

Det betyder, at LA's vælgere er mere liberale, når det gælder mulighederne for unges alkoholkøb end andre partiers vælgere. Men så er det også slut med velvilligheden, for så godt som ingen, heller ikke blandt LA's vælgere, ville slække på de daværende grænser.

Rygning

I undersøgelsen er der, som for alkohol, ingen forskel på rygevanerne mellem LA vælgere og andre partiers vælgere. Men i modsætning til holdningen til unges alkoholkøb er der heller ingen væsentlige forskelle på LA's og andre partiers vælgere.

Den nuværende grænse har været ved 18 år siden 2008, og halvdelen af vælgerne og endda også halvdelen af LA's vælgere svarer, at den grænse er for løs. Der er altså grænser for hvor liberale LA's vælgere er.

Er aldersgrænserne for unges køb af tobak…

	Stemte på Andet	*Stemte på LA*	*Total*
For lempelige / løse	702	47	749
	52.62%	50.54%	
Passende	608	44	652
	45.58%	47.31%	
For strenge	24	2	26
	1.80%	2.15%	
Total	1334	93	1427

Den nuværende grænse har været på 18 år siden 2008, og halvdelen af vælgerne og endda også halvdelen af LA's vælgere svarer, at den grænse er for løs. Der er altså grænser for hvor liberale LA's vælgere er.

Hvilke holdninger får vælgere til at stemme på Liberal Alliance

I ESS fra 2023 blev der stillet en række aktuelle holdningspørgsmål om hjemmearbejde, ukrainekrigen, covid, flygtninge, prisstigningerne og også de generelle ESS spørgsmål om tillid til statens autoriteter og tilfredshed med livet. De generelle spørgsmål blev stillet til alle respondenter, men de aktuelle holdningsspørgsmål blev stillet i batterier, hvoraf den enkelte respondent kun fik halvdelen af spørgsmålene i hvert batteri.

I det følgende udføres først en screening på det originale datasæt med manglende værdier af spørgsmålene for at finde hvilke spørgsmål, der kan bruges til at karakterisere LA's vælgere og dernæst udføres mere samlede analyser på et datasæt, hvor de manglende værdier er imputeret.

Screening af enkeltvariable

Ved hjælp af en SAS-makro blev det undersøgt om de enkelte holdningsspørgsmåls gennemsnit for LA's vælgere var signifikant forskelligt fra gennemsnittet for vælgere, der stemte på andre partier – bemærk at denne gruppe omfatter alle andre partier fra venstre til højre, så den er meget inhomogen. Makroen kaldte blot Proc Ttest i SAS med en class statement for hver af de mange holdningsspørgsmål. Det viste sig, at 35 spørgsmål gav signifikante forskelle med *p*-værdier mindre end 5%; de 10 mest signifikante er vist i næste tabel.

Variabel	*Obs*	*Differens*	*p*
Jeg er bekymret for, at mange mennesker vil glemme de vigtige læreprocesser fra pandemien og vende tilbage til tidligere vaner	691	1.93488	<.0001
Har/havde du mulighed for hjemmearbejdsdage på dit nuværende eller seneste arbejde?	1427	0.19435	0.0002
At agere bæredygtigt er en vigtig del af, hvem jeg er	718	1.37456	0.0002
Hvor stor tillid har du til politikere?	1421	0.87306	0.0006
Jeg føler, at min tillid til myndighederne er blevet påvirket positivt af deres håndtering af pandemien	705	1.40931	0.0007
Jeg føler, at min tillid til myndighederne er blevet påvirket negativt af deres håndtering af pandemien	684	−1.49623	0.0007
Jeg synes, at flygtninge bør have adgang til midlertidig eller permanent opholdstilladelse i mit land	666	1.56321	0.0008
Jeg er bekymret for, at Ukrainekrigen kan eskalere til en større international konflikt	677	1.34413	0.0010
Jeg mener, at individuelle handlinger ikke vil have en betydelig indvirkning på klimaforandringerne, medmindre store virksomheder også ændrer deres praksis	689	−1.35787	0.0015
Danmark bør være et foregangsland, hvad angår bæredygtig udvikling	685	1.44162	0.0015

Desuden blev det undersøgt med anden SAS-makro om de enkelte holdningsspørgsmål kunne forklare, at en vælger havde stemt på LA i en logistisk regressionsmodel for alle vælgere med LA som 'event'. Her blev resultaterne undersøgt ud fra AUC, 'Area Under the Curve', dvs. arealet under ROC kurven. De 10 mest betydningsfulde variable med højest AUC er vist i følgende to tabeller.

Variabel	*Odds Ratio*	*AUC*
Jeg er bekymret for, at mange mennesker vil glemme de vigtige læreprocesser fra pandemien og vende tilbage til tidligere vaner	0.763	0.685
Jeg synes, at flygtninge bør have adgang til midlertidig eller permanent opholdstilladelse i mit land	0.842	0.650
Jeg mener, at individuelle handlinger ikke vil have en betydelig indvirkning på klimaforandringerne, medmindre store virksomheder også ændrer deres praksis	1.231	0.649
Jeg er bekymret for, at Ukrainekrigen kan eskalere til en større international konflikt	0.814	0.648
At agere bæredygtigt er en vigtig del af, hvem jeg er	0.814	0.646
Jeg føler, at min tillid til myndighederne er blevet påvirket positivt af deres håndtering af pandemien	0.844	0.641
Jeg føler, at min tillid til myndighederne er blevet påvirket negativt af deres håndtering af pandemien	1.167	0.630
Danmark bør være et foregangsland, hvad angår bæredygtig udvikling	0.835	0.630
Jeg mener, at det er vores moralske pligt at hjælpe flygtninge, uanset omkostningerne	0.867	0.628
Jeg mener, at flygtninges ankomst kan være en byrde for landets økonomi	1.215	0.627

En samlet model for holdninger hos Liberal Alliances vælgere

Når alle holdningsvariable inddrages på en gang i en logistisk regressionsanalyse, kan modellen med alle variable naturligvis ikke estimeres, men der kan foretages en variabeludvælgelse.

I første omgang foretages en forward udvælgelse, hvor de mest betydningsfulde variable successivt tilføjes modellen indtil ikke flere bidrager signifikant målt på signifikans af parameteren. Der anvendes kun de variable, der i tabellerne over de variables enkeltvise betydning er enten signifikante i en parvis sammenligning med Proc Ttest på et 5% niveau eller har en AUC værdi over 0.6%

Da de fleste holdningsspørgsmål kun er besvaret af halvdelen af respondenterne, løber denne metode hurtigt tør for komplette observationer af alle de brugte variable i modellen. Derfor anvendes et imputeret datasæt, se indlægget ved symposiet januar 2024, der er komplet, men naturligvis præget af en vis usikkerhed. Især skal det vurderes ud fra, at de manglende værdier er beregnet ud fra de observerede værdier af andre variable, så der må forventes stærkere samvariation mellem variablene end der måske ville være i et oprindeligt fuldstændigt datasæt; altså en øget risiko for multikollinearitet.

I den variabeludvælgelse bliver 6 variable signifikante i den afsluttende model, som giver følgende estimater

Estimat	*StdErr*	*Wald X^2*	*Pr > ChiSq*	*Label*
0.3078	0.6528	0.2223	0.6373	*Intercept*
−0.1859	0.0483	14.8310	0.0001	*Hvor stor tillid har du til politikere?*
−0.0792	0.0329	5.7950	0.0161	*Danmark bør være et foregangsland, hvad angår bæredygtig udvikling*
−0.7048	0.2248	9.8303	0.0017	*Har/havde du mulighed for hjemmearbejdsdage på dit nuværende eller seneste arbejde?*
−0.0736	0.0330	4.9595	0.0259	*Jeg er bekymret for, at Ukrainekrigen kan eskalere til en større international konflikt*
0.1198	0.0579	4.2866	0.0384	*Jeg tror, at inflationen er midlertidig og vil aftage i de kommende år*
−0.1074	0.0526	4.1757	0.0410	*Jeg mener, at inflationen rammer lavindkomstgrupper hårdere end højindkomstgrupper*

De forklarende variable er alle angivet på en skala fra nul til ti, hvor nul er helt uenig og ti er fuldstændig enig. Det betyder, at LA's vælgere:

Ikke har tillid til politikere

Ikke har/havde mulighed for hjemmearbejde

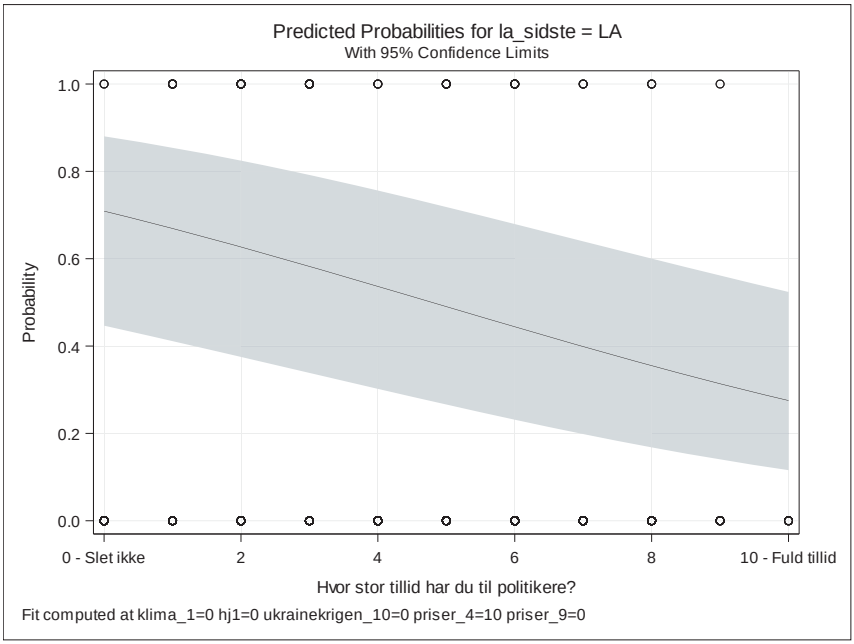
Ikke prioriterer bæredygtighed

Ikke er bekymret for at ukrainekrigen udvikler sig

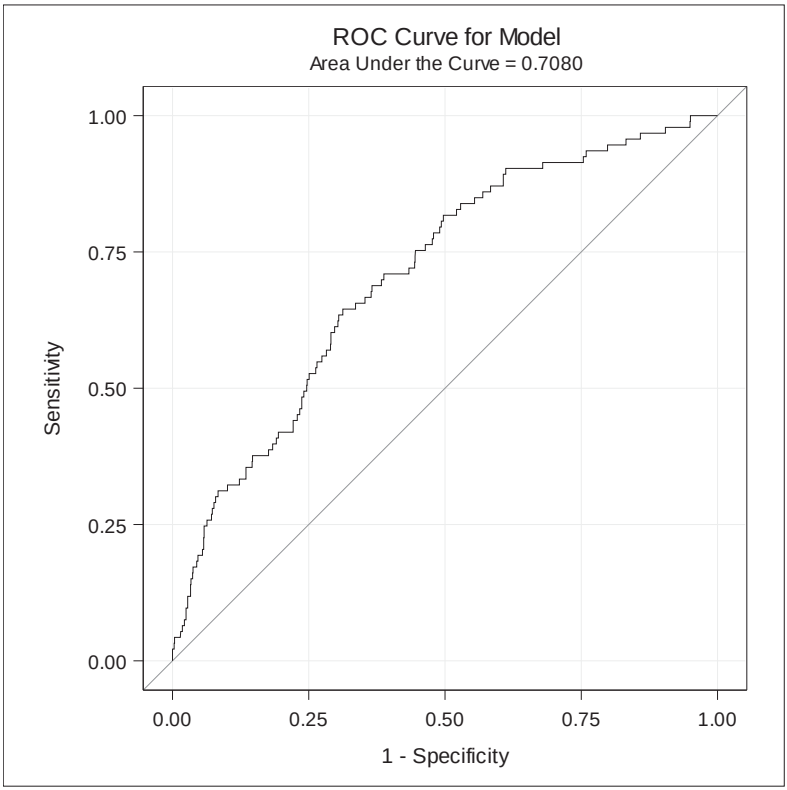
Tror at inflationen aftager snart

Ikke mener, at inflationen rammer de fattige hårdere end de rige

For at give et billede af, hvilke sandsynligheder, der er tale om, viser næste figur et effektplot, hvor effekten af variablen tillid til politikere varieres fra nul til ti, mens alle de andre variable fastholdes til den mest ekstreme værdi (nul eller ti), der fører til højest sandsynlighed for at stemme LA.



Modellen har en AUC på c = 0.708 beregnet ud fra ROC kurven



Kan BBR registeret bruges til retvisende ejendomsvurderinger?

Peter Linde, Peter@Brede.dk

Tidligere metodechef i Danmarks Statistik
Nævnsmedlem til vurderingsankenævnet 2023-2027
Bestyrelsen i Dansk Selskab for Surveyforskning og Symposium i Anvendt Statistik.
Censor på universiteterne i økonomi og statistik
Cand. scient. i matematisk statistik fra Københavns Universitet

Sammenfatning

I Skatteøkonomisk Redegørelsen 2021 står: *Det hidtidige ejendomsvurderingssystem har vist sig udfordret. Grund- og ejendomsvurderingerne har generelt været for upræcise, uensartede og uigennemskuelige og er desuden blevet kritiseret af Rigsrevision.*

Om de nye ejendomsvurderinger står der samtidig i redegørelsen: *De nye vurderinger af grund- og ejendomsværdier for ejerboliger er baseret på et bedre datagrundlag og en bedre statistisk model end det gamle vurderingssystem. Det giver mere præcise, ensartede og gennemskuelige vurderinger.* Står der.

I "Skatteøkonomisk Redegørelse 2021" står der alligevel på side 172: *Der kan også være udfordringer forbundet med kvaliteten af de data, som fremgår af Bygning- og Boligregisteret (BBR).* Denne afgørende forudsætning for retvise vurderinger er dog ikke blevet undersøgt nærmere i redegørelsen.

De fleste gennembearbejde modeller til estimation af ejendomsvurderinger vil virke, hvis data om boligerne er retvisende. Sikkert også den model, der er brugt til de foreløbige vurderinger. Som svar på et ønske om aktindsigt heri er der givet adgang til et notat: *Vurderingsmodel for ejerboliger 2019* (Se "kilder" til sidst). Notatet indeholder en gennemgang af modellen, men deler ikke centrale oplysninger om fx graden af forklaret variation, signifikans af modellens forklarende faktorer eller alternative løsninger.

Udfordringen for vurderingerne er, at datakvaliteten i BBR ikke gør det muligt at lave sikre retvisende estimater. En del af vurderingerne ligger inden for tolerancegrænsen i lovgivningen på plus/minus 20%, dvs. et samlet spænd på 40%. Det betyder, at en ejendomsvurdering med en modelbaseret ejendomsvurdering på 5 millioner ikke ændres, hvis den vurderingsfaglige vurdering er indenfor 1 million fra den modelbaserede. Også selv om den er baseret på fejl i BBR. Indenfor tolerancegrænsen på plus/minus 20% vil der ligge en del vurderinger, hvor estimatet påvirkes reelt af forkerte BBR oplysninger.

En vurdering af en ejendom afhænger ikke af historiske salgspriser for ejendommen, men af de 15 nærmeste nabohandler. Selv om boligejeren har indberettet korrekt til BBR vil ejendomsvurderingen også afhænge af, om naboerne har indberettet korrekt til BBR.

Udfordringerne med de nye ejendomsvurderinger er ikke primært et IT problem eller den statistiske model - det er dataproblemet i selvrapporterede oplysninger til BBR, uden egeninteresse i de er retvisende. Man har i hele processen overset den største udfordring – at skabe et system til kvalitetsudvikling af BBR, selv om de røde lamper blinkede.

Indledning

De nye ejendomsvurderinger blev oprindeligt estimeret til at koste knap 100 millioner kroner at udvikle. Her knap 10 år efter beslutningen om at udarbejde ejendomsvurderinger baseret på BBR og statistiske modeller, er der søgt tillægsbevillinger i Finansudvalget på over 4 milliarder kroner. Og vi har stadig ikke en løsning, der virker stabilt.

De historiske salgspriser, geodata samt Bygnings- og Boligregisteret (BBR) er de væsentligste datakilder til statistiske regressionsmodel, der estimerer de nye og foreløbige ejendoms- og grundvurderinger. BBR blev oprettet i 70-erne og det er boligejernes ansvar, at alle oplysningerne om areal, opvarmning, tilbygninger mv. er retvisende. Udfordringen for kvaliteten af BBR er, at boligejerne ikke har nogen egeninteresse i, der står det rigtige i BBR. Hvis arealet er angivet for lille, tilbygninger eller forbedringer ikke er oplyst, vil det give lavere ejendomsvurderinger.

Den afgørende forudsætning for alle statistiske modeller er at data er retvisende. Ellers kan de ikke bruges til analyser og estimationer. Forudsætningen er også vigtig fordi i komplekse statistiske modeller og "kunstig intelligens" (Artificial Intelligence) kan IT udfordringer, fagudtryk og matematiske formler skygge for det grundlæggende: Kan data overhovedet bruges? Estimationen af ejendomsvurderingerne kan aldrig blive bedre end kvaliteten af input data, men vil ofte blive markant dårligere. Det skyldes at den statistiske model i sin optimering med forkerte data, kan give systematiske og vilde outliers. Specielt hvis der bruges komplekse modeller og ikke robuste metoder. At en statistisk model er kompleks og svær at forstå, er ikke i sig selv et kvalitetsstempel.

Ovenstående var baggrunden for jeg, som uddannet matematisk statistiker og tidligere metodechef i Danmarks Statistik, med ansvaret for udarbejdelse af varedeklarationer for blandet andet BBR, søgte om aktindsigt om:

1) De bagvedliggende statistiske regressionsanalyser (IA), der ligger bag modellen til fastlæggelse af ejendoms- og grundvurderingen.
2) De bagvedliggende bilag til punkt 1
3) Den endelige model og notater/redegørelser, der beskriver denne. Herunder betydningen sf generelle landsdækkende forhold/relationer, lokale og vægtningen af nærmeste handler.
4) Hvordan ovenstående er formidlet til ledelsen i vurderingsankestyrelsen samt til politikere/departementet og ministeren.

Ansvaret for BBR og retsgrundlaget

Det er kommunerne, der er registerfører for oplysninger i BBR og det er ejerne af boligerne, der er forpligtiget til der står de rigtige oplysninger om boligen de ejer. For at give kommunerne et retsgrundlag giver BBR loven grundlag for at give bøder på op til 5.000 kr. for urigtige oplysninger. Det er ikke lykkedes at finde et eneste tilfælde, hvor denne bestemmelse har været anvendt i 50 år, og der findes ikke retningslinjer for, hvordan kommunerne skal fastlægge bødestørrelsen mere konkret. BBR har løbende tilsyn af kommunernes opgaver med BBR og på BBR's hjemmeside findes 22 tilsyns-

rapporter[1] fra 2021-2024. Ikke i en eneste tilsynsrapport er det oplyst, der er givet bøder. Spørgsmålet behandlers ikke i opfølgningen eller konklusionen i tilsynsrapporterne.

Ved udbetalingen af støtte til gasopvarmning i 2022/2023 blev det synligt, at kvaliteten af BBR ikke er godt nok til retvisende myndighedsafgørelser. I 2022 besluttede Folketinget, at de omkring 405.000 husstande, der havde gasopvarmning ifølge BBR, skulle have et skattefrit beløb afhængig af indkomst. I alt blev der udbetalt 2,4 milliarder kroner. Mindst 147 millioner kroner blev udbetalt forkert. Det blev besluttet, at fejludbetalinger, baseret på der i BBR forkert stod, at ejendommen var gasopvarmet, ikke skulle tilbagebetales. Ifølge loven om BBR er det ejerens ansvar, at oplysningerne er retvisende, og mangler kan betyde en bøde på op til 5.000 kr. Med beslutningen om at fejlindberetninger til BBR ikke skulle tilbagebetaltes blev det modsatte retsprincip understøttet. Der blev ikke rejst bødekrav og fejludbetalingen kunne beholdes. Det udfordrer hvilke krav, der fremadrettet kan rejses overfor forkerte BBR indberetning.

Kvaliteten af registre og BBR

Grundlæggende afhænger kvaliteten af et register om at alle dem, der indberetter, bruger registeret eller er ansvarlige for registeret, har interesse i at oplysningerne er rigtige. Samt af hvor krævende (byrden) det er at indberette korrekt. Interessen for at indberette korrekt kan fx skyldes bødekrav eller økonomiske tilbagebetalinger. Godkendelse af selvangivelsen for skat er ligesom indberetninger til BBR baseret på selv-angivelser. BBR er primært en indberetning på eget initiativ, og skatteopgørelsen modsat en godkendelse af registrerede oplysninger og kontrol fra flere kilder. En lønmodtager og arbejdsgiver kan fx have fælles interesse i at lønnen oplyses korrekt og kontrollerer på den måde hinanden. Så denne oplysning er som udgangspunkt mere retvisende end fx oplysningen om arbejdstid og stilling, der også følger med arbejdsgiverens indberetning om løn. Tilsvarende har skolerne en interesse i at indberette nye elever til elevbestanden, fordi det giver øget støtte. Denne oplysning er som udgangspunkt af en bedre kvalitet, end indberetninger om ophørte elever i studieforløbet. Det samme gælder indberetningen af diagnose og sygdom, der efterfølgende er behandlet eller viste sig ikke rigtige.

I BBR er udfordringen, at ejeren ikke har en økonomisk fordel af at indberette øgede boligareal, nye toiletter eller udbygninger, da det øger ejendomsbeskatningen. Køber har heller ikke en interesse i at sælger oplyser fx øget areal, da de så får øget ejendomsskat. Hvis den nye ejer omvendt bagefter opdager, at arealet er for højt i BBR, er der højesteretsafgørelser, der understøtter, der kan rejses krav mod sælger og den nye ejer vil rette det i BBR, da det betyder mindre beskatning. Så når arealet ikke er korrekt i BBR, er det oftere for lavt end for højt.

[1] Bornholm, Skive, Favrskov, Glostrup, Lolland, Frederikshavn, Faarborg-Midtfyn, Odsherred, Ikast-Brande, Jammerbugt, Ringkøbing-Skjern, Odense, Vordingborg, Frederikssund, Sønderborg, Slagelse, Nordfyn, Greve, Stevns, Hjørring, Nyborg, Varde og Holstebro.

Om kvaliteten i BBR skriver BBR:

Det er et tilbagevendende spørgsmål, hvor god datakvaliteten i BBR faktisk er. Svaret er tæt knyttet til spørgsmålene om, hvad det er man ønsker at registrere i BBR og hvad det skal anvendes til. Da formålet med BBR er mangesidigt bliver kvalitetsspørgsmålet det også. Derfor kan der ikke gives noget eksakt og generelt svar på, hvor god datakvaliteten er, da det afhænger af den sammenhæng, BBR-data skal indgå i[2].

BBR-myndigheden forsøger alligevel at finde metoder til at måle datakvalitet. Den seneste måling af datakvaliteten i BBR er foretaget af SAS Institute, som ifm. arbejdet i regeringens ekspertudvalg om ejendomsvurderingen, gennemførte en omfattende analyse af BBR-data i 2014.

Man skal dog, som det fremgår af rapporten fra SAS Institute, være opmærksom på, at den anvendte metode med såkaldte regelscanninger, udmærker sig ved at finde observationer, der enten er objektivt forkerte eller afviger fra normalen for de undersøgte grupper. Fejlbehæftede observationer, der ligger inden for et plausibelt interval, er langt vanskeligere at opdage, og vil typisk ikke kunne identificeres ved regelscanninger af denne karakter.[3]

Hvis sådanne fejl skal måles, er en grundig manuel stikprøvekontrol med besøg på ejendomme en mulighed for at af- eller bekræfte kvaliteten af data. Udfordringen ved en sådan stikprøve er, at den skal afspejle den høje kompleksitet, som kendetegner den danske boligmasse, hvor der findes mange særtilfælde inden for hver boligtype. Et betydeligt antal manuelle tjek ville derfor være en nødvendighed.[4]

Det er således ikke umiddelbart muligt hos BBR at finde oplysninger om BBR registeret har en kvalitet, der understøtter, det er muligt at bruge til retvisende ejendomsvurderinger. Det skal undersøges konkret da det "afhænger af den sammenhæng BBR-data indgår i." En sådan undersøgelse med manuelle tjek er ikke lavet i forbindelse med brugen af BBR som primærkilden til de nye ejendomsvurderinger.

Danmarks Statistik bruger BBR registeret til statistik og skriver følgende i varedeklarationen om statistikken:

Det er i høj grad op til boligejere selv at opdatere oplysninger i BBR, derfor kan man tvivle på at ændringer fx i antal værelser, toilet-/badeforhold eller skift af opvarmning altid bliver meddelt BBR.[5]

Hovedkilde til usikkerhed er, at der er mangler i rapporteringen. Manglerne skyldes uvidenhed, forsømmelighed, glemsomhed eller andet. Det er for eksempel usikkert i hvor stort omfang, borgeren melder ændrede antal værelser når der laves en ombygning eller udskiftning af varmekilde.

[2] Forfatters understregning. Her er det myndighedsafgørelser om beskatning.
[3] Forfatters understregning
[4] Forfatters understregning
[5] Forfatters understregning

Danmarks Statistik tager på denne måde et væsentligt og generelt forbehold overfor sikkerheden af oplysningerne.

Når man søger i sager på internettet, bekræftes at BBR er udfordret på en lang række punkter. DR har spurgt beskikkede landinspektører om de er enige i, at man kan regne med oplysningerne i BBR-registeret er korrekte. 6% er enige i de er rigtige, mens 58% er uenige eller meget uenige i de er rigtige. Til DR siger en landinspektør om BBR-oplysningerne: *Jeg bliver overrasket når det passer.* LIFA, der arbejder med opmåling af ejendomme, skriver fx: *Det er alment kendt at en meget stor del af arealerne i BBR er fejlbehæftede*. LIFA vurderer, at der er fejl i op til 20% af de registrerede boligarealer. Konkrete tjek hos LIFA har vist at 20-25% var forkerte.

Selv om der har været mange advarsler og kendskab til at BBR ikke havde en særlig høj kvalitet er der ikke udviklet et egentligt driftssystem til at sikre den løbende kvalitet. Mere brugervenlige indberetninger løser det ikke. Mange data er ikke det samme som kvalitet. Egentligt kunne man stoppe artiklen her, for der er ikke grundlag for at antage, at BBR registeret kan bruges til retvisende ejendomsvurderinger, uden betydeligt understøttende vurderingsfaglig sagsbehandling og et nyt driftssystem. For fuldstændighedens er der alligevel et kort overblik over modellen nedenfor.

Estimationsmodellen

Den statistiske model er nærmere beskrevet i bilaget **Vurderingsmodel for ejerboliger** 2019, Implementeringscenteret for Ejendomsvurderinger, Udviklings- og Forenklingsstyrelsen, Modelkontoret, der findes på www.statistiksymposium.dk under Symposiebogen 2025.

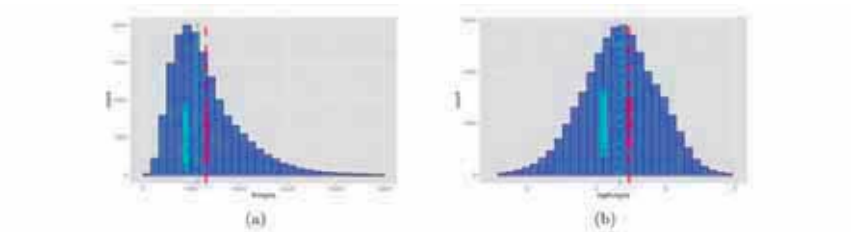


Figure 1: *Transformation fra kvadratmeterpriser(a) til log-kvadratmeterpriser(b) viser at transformationen reducerer skævheden og fordelningen bliver normalfordelt, hvilket er en forudsætning for middelrette estimatorer i en statistisk model.*

Modellen for ejendomsværdien bygger på **kvadratmeterprisen**, som er handelsprisen divideret med det sammenvejede boligareal, som det er kendt fra salgsopstillinger fra ejendomsmæglere. Handelsprisen reguleres med tidsudviklingen, så de kan sammenlignes. Kvadratmeterprisen, der er responsvariablen i den efterfølgende regressionsanalyse, er højreskæv. Det bruges i notatet som argument for at transformere

kvadratmeterprisen med logaritmen. Den logaritmiske fordeling af kvadratmeterpriserne er mere normalfordelt. Begrundelsen for transformationen er ifølge notatet, at det er en forudsætning for middelrette estimater i en statistisk model. Transformationen har samtidigt den konsekvens, at de forklarende variabler og faktorer påvirker kvadratmeterprisen multiplikativt. Dvs. en resulterende faktor 1,1 for fx WC, påvirker kvadratmeterprisen med 10%, frem for fx en addidativ virkning på fx 1.000 kr. for et WC, uanset hvor høj kvadratmeterprisen er. Uanset om det er en multiplikativ eller adidativ virkning, ganges kvadratmeterprisen til sidst med det sammenvejede boligareal for at for at estimere ejendomsværdien, så WC-et i en stor ejendom tæller mere, uanset addidativ eller multiplikativ virkning i modellen. Det giver god mening, da der normalt også er større og bedre WC-er i store end i små ejendomme. Det ses generelt i modeller om ejendomsvurderinger, at en multiplikativ virkning giver en pæn 'grad af forklaret variation' og bedre modelfit. Men det behøver ikke at være tilfældet og er ikke beskrevet i notatet om den konkrete analyse. Her begrundes den logaritmiske transformation med at ejendomspriserne hermed bliver normalfordelte. Normalt ville man statistikfagligt tage udgangspunkt i om afvigelserne (residualerne) er normaltfordelte. Dette er ikke undersøgt.

Det første trin er en **naboprismodel**, hvor handler i nærheden af den ejendom, der vurderes, får højst værdi, når de ligger tæt ved ejendommen og mindst, når de ligger længere fra. Vægtene estimeres og bestemmes vha. en kernefunktion og summer til 100%. Det samvejede gennemsnit beregnes som:

$$\ln(\hat{p}_i)_K = \sum_{n=1}^{N} \ln(p_n)\omega(d_n)$$

Gennemsnit af N observationer, der er responsvariablen i regressionen, er ifølge den centrale grænseværdisætning i sig selv asymptotisk normalt fordelt. Uanset transformation af logaritmen.

I et vist eksempel i notatet med 15 nabohandler er den højeste vægt 15% og den mindst 1%. Det betyder de nærmeste 4 handler tæller for godt halvdelen. I hvilken grad det giver større præcision eller mere variation for den enkelte vurdering er ikke beskrevet, men vurderingen bliver i alle tilfælde mere afhængig af om BBR oplysningerne for de allernærmeste handler er indberettet korrekt. Hvis moderniseringer eller det fulde areal ikke er indberettet korrekt i de allernærmeste handler, har det stor og følsom betydning for at kvadratmeterprisen stiger, selv om den ejendom, der vurderes, har indberettet korrekt.

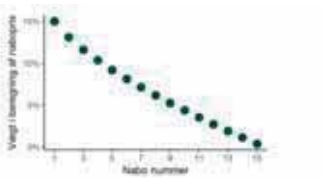


Figure 2: Eksempel på vægte ved 15 nabober.

Vægtene rangordens efter afstand, og hvis der er 15 nabohandler, er det de samme låste procenttal man bruger. De fx 15 handler findes ud fra en cirkel med centrum i den bolig, der estimeres. I notatet begrundes dette med, at "metoden svarer til det en ejendomsmægler ville gøre i en vurderingssituation. Metoden er således valgt ud fra et vurderingsfagligt synspunkt." I den ejendomsvurdering jeg selv har fået tilsendt, var en cirkel dog ikke den vurderingsfaglige rigtige tilgang. Af de 15 handler var 5 (33%) i min rækkehusbebyggelse på 188 rækkehuse, hvor der en 1. sal over en del af huset og lodrette vægge. De andre 10 (67%) handler er fra et naboområde med en helt anden type rækkehuse med kælder og 1. sal med skrå vægge. Her ville en lokal ejendomsmægler sammenligne med handler fra samme boligforening. Det er det der sker, når der ikke forud for modelarbejdet er indhentet lokale oplysninger om fx sammenhængende bebyggelser og ejerlaug for at kvalitetssikre vurderingen. Det betyder for en konkret ens bebyggelse - at huse i midten af bebyggelsen, sammenlignes med andre huse i bebyggelsen, medens huse i randen af bebyggelser i væsentlig grad sammenlignes med huse fra andre bebyggelser. Oplysninger om ejerlaug er offentlig tilgængelige oplysninger, som kommunerne kan hjælpe med. Min bebyggelse på 188 rækkehuse udgør 1% af kommunes indbyggere, og kommunen 1% af landets indbyggere. Lokal fundering af modellen har ikke været en del af udviklingsbudget på over 4 milliarder.

Den Hedoniske regressionsmodel for logaritmen af kvadratmeterprisen for bolig nummer "i" opstilles som:

$$\chi \equiv \{x_p, p = 1, \cdots, N\} \tag{3}$$

som komplet set af ejendommens karakteristika. Regressionsmodellen har specifikationen:

$$\ln(\hat{p}_i)_G = \hat{\beta}_0 + \sum_{j=1}^{J} \hat{\beta}_j x_{i_j} + \sum_{h=1}^{H} s_h(x_{i_h}) + \epsilon_i, \qquad J + H = N \tag{4}$$

$\hat{\beta}_0$ er det estimerede konstantled, mens $\hat{\beta}_j$, $j = \{1, \cdots, J\}$, er de estimerede lineære koefficienter for værdien af ejendom i's j'te karakteristika, herunder dummy-variable for kategoriske variable.

$s_h(x_{i_h})$ beskriver de karakteristika, der beskrives ved hjælp af ikke-parametriske spline-funktioner, $s_h(\cdot)$.

Parameteren β beskriver de J procentvise tillæg til ejendommens kvadratmeterpris. I notatet står der, at "det giver letfortolkelige parametre, der kan anvendes til at kommunikere værdifastsættelsen af ejendommen i forhold til ejendomsejer og vurderingsmedarbejdere". I den ret omfattende skrivelse på op mod 50 sider man modtager, når man får en ejendomsvurdering, er procenttillæggene dog ikke gengivet. Det betyder vurderingen kan fremstå som en "sort boks".

Bias i modellen

For ejendomme med gennemførte handler er det muligt at estimere afvigelsen mellem salgsprisen og den estimere salgspris. I notatet er redegjort for, hvordan dette residual kan opdeles i tre komponenter:

$$\xi_i = \ln(\widehat{p_{g,i}}) - \sum_{\substack{n=1 \\ n \neq i}}^{N} \ln(\hat{p}_{g,n}) \cdot \omega(d_n) = \zeta_{i2} + \zeta_{i3} + \zeta_{i4}$$

$$= \left(\sum_{j=1}^{J} \hat{\beta}_{i_j} x_{i_j} - \sum_{\substack{n=1 \\ n \neq i}}^{N} \left[\sum_{j=1}^{J} \hat{\beta}_{n_j} x_{n_j} \cdot \omega(d_n) \right] \right) + \left(\sum_{h=1}^{H} s_{i_h}(x_{i_h}) - \sum_{\substack{n=1 \\ n \neq i}}^{N} \left[\sum_{h=1}^{H} s_{n_h}(x_{n_h}) \cdot \omega(d_n) \right] \right)$$

$$+ \left(\epsilon_i - \sum_{\substack{n=1 \\ n \neq i}}^{N} \epsilon_n \cdot \omega(d_n) \right) \tag{16}$$

For disse gælder at ζ_{i2} og ζ_{i3} kan beregnes via regressionsmodellen, mens den del af ζ_{i4}, der kan tilskrives nabosalgenes statistiske usikkerhed $\sum_{\substack{n=1 \\ n \neq i}}^{N} \epsilon_n \cdot \omega(d_n)$ er givet idet salgene kan observeres.

Der er to kvalitetsmål i notatet - det ene er gengivet nedenfor. Det viser, der er størst sikkerhed i større byer. Der må således forventes flest klager for områder med de lyse farver i figuren nedenfor, hvor over halvdelen af vurderingerne ligger uden for intervallet plus/minus 20% af salgsprisen.

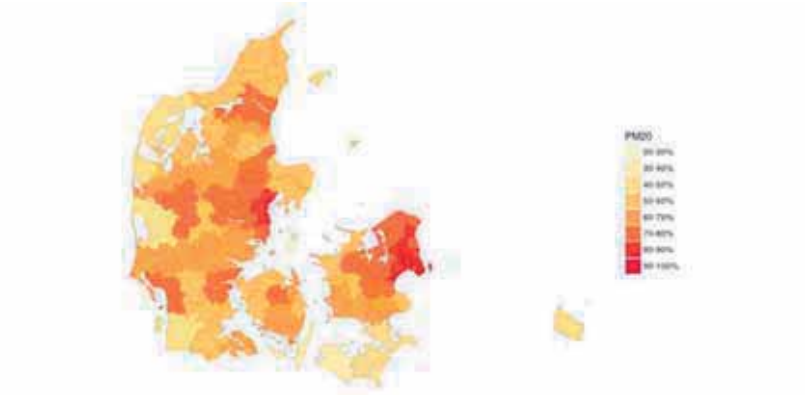


Figure 8: PM20 på kommuneniveau

Figur (8) viser andelen af boliger, der prædikteres indenfor ±20% af den reelle salgsværdi fordelt på landets 98 kommuner. I prækiktionen er anvendt den samlede model, og resultaterne er baseret på testsættets observationer.

Ejendomsværdiens opdeling i grund og bebyggelse

En særlig udfordring i **e**jendomsvurderingen (**E**V) er, at den skal opdeles i vurdering af **g**runden (**G**V) og **b**ebyggelsen (**B**V), fordi grundværdien beskattes separat med grundskylden til kommunen.

uafhængig køber. Fra Vurderingslovens §17 defineres at *Ved grundværdien forstås værdien af grunden i ubebygget stand under den forudsætning, at grunden vil blive anvendt og udnyttet bedst muligt i økonomisk henseende.* For at den logiske sammenhæng

$$EV = BV + GV \quad \Leftrightarrow \quad GV = EV - BV \tag{20}$$

kan fastholdes, defineres bygningsværdien BV således som værdien af alle de bygningsdele og anlæg, der findes på grunden. I forhold til en faktisk bebyggede ejendom, kan der være tale om at den bebyggelse, der er på grunden ikke nødvendigvis er *den bedst mulige i økonomisk henseende.*

Når en ejendom handles, er det en samlet pris, der ikke deles op i to priser/komponenter for bebyggelse og grund. Vurderingen af grundværdien søges løst ved at vurdere markedspriserne ved salg af ubebyggede grunde. Det siger selv, at datagrundlagt for vurdering af grundværdien (GV) bliver med et anderledes datagrundlag og grunde, der ligger længere væk, og dermed modelbaseret på en anden måde end ejendomsvurderingen. I notatet er behandlet, hvordan dette søges løst bedst muligt med grundværdikurven. Det fremgår af notatet, at grundværdierne kommer til at udgøre en større andel af ejendomsvurderingen. Det har også betydning for den usikkerhed, der opstår og har også givet en del debat.

Opsamling

I indledningen til artiklen blev det fremhævet, at udfordringen ved ejendomsvurderingerne ikke primært er et IT eller modelproblem, men et dataproblem ved pga. af kvaliteten af BBR. Modellerne og IT løsningerne udfordres heraf, og bliver budbringeren for at data ikke er retvisende. Men ikke årsagen.

Når der laves en konkret ejendomsvurdering, skal man huske, at det både er et spørgsmål om alle **forbedringer er indberettet til BBR** og den **statistiske afvigelse**. De ejendomme, der ikke har oplyst alle forbedringer til BBR vil få mindre ejendomsvurderinger, og de ejendomme, der har rigtige oplysninger højere vurderinger. Specielt hvis det er tilfældet for de nærmeste ejendomme. Det betyder for disse ejendomme, at de bliver pålagt højere ejendomsskatter ved salg. Den statistiske afvigelse, som er den der er beskrevet i figuren med danmarkskortet ovenfor, er et nul-sums spil. Det betyder, at der generelt for hver ejendom med en for høj vurdering er en med en tilsvarende lavere. De to komponenter – BBR fejlen og den statistiske fejl er de væsentligste man møder ved vurderingerne.

Hvis man ville løse udfordringen med kvaliteten af BBR kræver det betydelige ressourcer i kommuner og på landsplan i driften samt at lovgivning om bøder efterleves. Men det kræver nok også en ny lovgivning og retspraksis. Man kunne fx forestille sig følgende:

1) Ejere får frit lejde for ændringer i BBR indtil en given skæringsdato, fx 1. januar 2026. Ændringer der ville resulterer i en anden (højere) vurdering, har ikke konsekvenser og der gives ikke bøde.
2) Efter skæringsdatoen vil ændringer ikke kun blive beskattet fra skæringsdatoen, men også pålagt et tillæg på fx 25%.

3) Når boligen sælges, skriver både sælger og køber under på at BBR oplysningen er korrekt, som en forudsætning for tinglysning af handlen. Og køber overtager sælges ansvar fra skæringsdatoen. Køber kan tegne en forsikring herfor eller bruge konsulenthjælp hertil.
4) Hvis tillæg efter skæringsdatoen ikke kan indfris, indefryses det som gæld i ejendommen med en løbende forrentning.

En ordning som ovenfor vil motivere **både** sælger og køber til at sikre korrekte oplysninger i BBR. Den nødvendige kvalitet af BBR er en forudsætning for, der kan træffes retvisende myndighedsafgørelser om ejendomsvurderinger på baggrund af BBR. Som nævnt tidligere skal der i det hele taget sikres et system til løbende udvikling af BBR, hvis det fremadrettet skal bruges til myndighedsopgaver, ligesom man gjorde med selvangivelserne af skat.

Som afslutning citatet fra Skatteministeriets Skatteøkonomisk Redegørelsen 2021: ***Der kan også være udfordringer forbundet med kvaliteten af de data, som fremgår af Bygning- og Boligregisteret (BBR).*** Ukendt var det ikke!

Ingen kæde er stærkere end sit svageste led.

Kilder

Spørgsmål 1-3 i ansøgningen om aktindsigt.
Vurderingsmodel for ejerboliger 2019, Implementeringscenteret for Ejendomsvurderinger, Udviklings- og Forenklingsstyrelsen, Modelkontoret, der findes på www.statistiksymposium.dk i Symposiebogen 2025 som bilag.

Spørgsmål 4 i ansøgningen om aktindsigt
Skatteøkonomisk redegørelse fra 2021 kapitel 5 (side 163-197) som findes på skatteministeriets hjemmeside. https://skm.dk/aktuelt/publikationer/rapporter/skatteoekonomisk-redegoerelse-2021

Arbejdsmarkedsreformer – øger de beskæftigelsen?[1]

Professor, dr. scient. adm. Jesper Jespersen

Roskilde Universitet

Indlæg på konference i Anvendt Statistik, den 27.-28. januar 2025, Odense

Abstrakt

Arbejdsmarkedsreformer har stået stærkt i den økonomiske politik fra slutningen af 1970erne. Men med helt forskellig prioritering i afvejning mellem udbud af og efterspørgsel efter arbejdskraft. I perioden frem til midten af 1990erne lå fokus på en reduktion af arbejdsudbuddet, bl.a. gennem en reduktion af den ugentlige arbejdstid, øget ferie og (tvungen) afspadsering af overarbejde i kombination med en efterspørgselsorienteret økonomisk politik.

Grundlaget for arbejdsmarkedspolitikken blev efterfølgende ændret radikalt gennem Velfærdskommissionens (2003-06) redegørelser. Her anbefaledes det at øge arbejdsudbuddet gennem mindskede sociale ydelser, reduceret indkomstskat samt øget pensionsalder, hvilket ville ifølge den neoklassiske økonomisk teori føre til en tilsvarende forøgelse af beskæftigelse.

Artiklen analyserer primært udviklingen på det danske arbejdsmarked igennem de seneste 15 år med fokus på den nyere og mere detaljerede statistik. Med dette afsæt sættes fokus på det spørgsmål, om det er striben af udbudsreformer, der siden 2006 har kendetegnet den økonomiske politik, eller om det er efterspørgslen efter arbejdskraft, der er den dominerende faktor bag den markant U-formede udvikling i både arbejdsudbud og beskæftigelse igennem periode?

Indledning

I perioden efter 2001 har arbejdsmarkeds- og socialreformer med sigte på at øge arbejdsudbuddet og dermed – ifølge neoklassisk teori - beskæftigelsen nærmest stået i kø som de skiftende regeringers højest prioriterede initiativer, uanset regeringens farve. Reformerne synes stadig ingen ende at ville tage, hvilket i praksis har betydet bl.a. øget

[1] Dette er en omarbejdet og opdateret artikel i forhold til Jespersen, J. (2023)

pensionsalder, mindskede (arbejdsmarkedsrelaterede) sociale ydelser og en reduktion af progressionen i indkomstskatteskalaen.

Som det vil fremgå af denne artikel, har resultaterne af de mange reformer mildt sagt været flertydige, idet både arbejdsudbud og beskæftigelse i lange perioder har ligget væsentligt under det niveau, som kommissionsøkonomerne havde stillet politikerne i udsigt.

Helt tilbage til 'Velfærdskommissionens' rapport i 2006[2]var anbefalingen fra de deltagende (overvejende neoklassiske) økonomer, at både en forøgelse af pensionsalderen, beskæring af efterlønnen, reduktion af sociale ydelser og mindsket progression i indkomstskatteskalaen ville udgøre vigtige instrumenter for at øge arbejdsudbud og dermed beskæftigelsen.

Men sådan gik det ikke umiddelbart. Ikke mindst viste der sig at være en afgørende forskel på, om det var tilbagetrækningsalderen, der blev hævet, eller om det var dagpenge-, kontanthjælpsniveauet og/eller indkomstskatten for de beskæftigede, der blev sænket.

Den helt afgørende forøgelse i beskæftigelsen indtrådte tilsyneladende først, da den økonomiske politik blev lagt om i kølvandet på udbruddet af Covid19-epidemien. Her blev Budgetloven suspenderet, og en efterspørgselsorienteret økonomiske politik gennemført (ikke kun i Danmark, men i EU generelt). Det fik beskæftigelsen opgjort som antallet af personer (og arbejdstimer) til at overstige niveauet fra 2008 og nåede sit foreløbige højdepunkt i 2023. I hvilket omfang de mange arbejdsmarkedsreformer har bidraget til denne gunstige udvikling, mangler endnu at blive endeligt klarlagt.

Denne artikel vil især fokusere på de teoretiske konklusioner, som striben af arbejdsmarkedsfokuserede kommissioner er nået frem til igennem de seneste 20 år og sammenholde dem med virkeligheden. En oplagt forklaring på disse markante afvigelser synes at være, at anbefalingerne alle direkte eller indirekte lænede sig op af en teoretisk ramme baseret på en generel ligevægtsmodel for dansk økonomi udviklet af velfærdskommissionen under navnet DREAM (Danish Rational Expectation Annual Model), Velfærdskommissionen (2004). Den er efterfølgende blevet benyttet af flere ministerier ved kvantificering af arbejds*udbud*sorienterede politiske forslag, se f.eks. skatteministeriet, (2012). Den i finansministeriet benyttede regnemodel, ADAM udviklet af Danmarks Statistik, blev efterfølgende tunet, så de langsigtede resultater – fuld (strukturel) beskæftigelse - flugtede med DREAM-modellen. Dog med den forskel at det i ADAM-modellen tager adskillige år fra en udbudsreform vedtages og efterfølgende implementeres til den fulde beskæftigelseseffekt er opnået, Finansministeriet, 2024.

[2]file:///C:/Users/jesperj/Downloads/Velf%C3%A6rdskommissionen+%E2%80%93+form%C3%A5l,+resultater+og+erfaringer.pdf

Arbejdsmarkedet: hvordan beskrives det statistisk?

Denne artikel må dog indledes med noget så kedeligt som en beskrivelse af de statistiske udfordringer, som arbejdet med den talmæssige belysning af arbejdsmarkedet indebærer. Uanset om vi taler om beskæftigelse, arbejdsstyrke eller ledighed, så er der varierende statistiske definitioner og opgørelsesprincipper, der hver især leder til forskellige empiriske resultater.

Et klassisk eksempel på en sådan forskel er den statistiske opgørelse af arbejdsløsheden, der alene i Danmark kan variere med op til 100.000 personer på den samme population og tidsperiode. Hertil burde der lægges et skøn over den (statistisk) skjulte arbejdsløshed, som der ifølge sagens natur ikke er (præcise) tal for. De tre af Danmarks Statistik regelmæssigt offentliggjort tal for arbejdsløsheden lød i 2022 på hhv. 70.000 personer, 90.000 personer (begge registerbaserede) og 140.000 personer (Arbejdskraftundersøgelse (AKU) baseret på interviews, der følger EuroStat's anvisninger)[3]. Alt mens medierne (og delvis) politikerne talte om udbredt mangel på arbejdskraft.

Hvordan hænger det sammen? – og kan alle have ret? Det korte svar er desværre et tøvende 'ja', fordi der benyttes (og delvis argumenteres) ud fra ret så forskellige opgørelsesprincipper og hensyn, alt efter hvilket spørgsmål, der skal belyses/besvares. De registerbaserede tal er en opgørelse af det antal (omregnet til fuldtid) personer, der modtager arbejdsmarkedsydelser fra offentlige myndigheder (Dagpenge, kontanthjælp, beskæftigelsestilskud mm.). Formålet er her primært at belyse belastningen af de offentlige budgetter. Da f.eks. dagpengeperioden blev reduceret fra 4 til 2 år, faldt den registrerede arbejdsløshed – derimod stort set ikke den interviewbaserede arbejdsløshed. De mange ændringer i de arbejdsmarkedsrelaterede ydelser og dermed den statistiske afgrænsning har ændret opgørelsesprincipperne i den officielt (mest) benyttede arbejdsløshedsstatistik, hvilket i sig selv er en udfordring, der delvis kunne afhjælpes ved brug af tal fra AKU-statistikken, der er baseret på telefoninterview.

Tilsvarende problemer melder sig imidlertid også delvist ved brug af den interviewbaserede statistik. Idet ændres spørgsmålene, så ændres svarene sig naturligvis også. Hvornår er man arbejdsmarkedsparat? Oplever man sig selv som arbejdsmarkedsparat (en del af arbejdsudbuddet), når man igennem 10 år har stået bagest i arbejdsløshedskøen? Her nærmer vi os begrebet 'skjult arbejdsløshed'. Konstateres må det i hvert fald, at både i perioden 2005-08 og nu igen 2021-23 er beskæftigelsen steget betydeligt mere, end det skulle være muligt ud fra de statistiske oplysninger, der forelå om arbejdsudbuddets omfang lige inden opsvinget tog fart - uanset hvilken af statistikkerne, der benyttes! Da lavkonjunkturen i 2009 satte ind, faldt beskæftigelsen med ca. 200.000 personer; mens den registerbaserede arbejdsløshed kun steg med godt 100.000 personer. Årsagen hertil er flere;

[3] https://www.dst.dk/da/Statistik/emner/arbejde-og-indkomst/befolkningens-arbejdsmarkedsstatus/arbejdsmarkedsregnskab

men en væsentlig forklaring er, at adskillige personer ikke var berettiget til dagpenge- eller kontanthjælp og derfor forsvandt ud af den officielle arbejdsmarkedsstatistik.

Det er således helt afgørende hvilket statistisk grundlag, der benyttes ved analysen af ændringer i arbejdsløshed og dermed af det potentielle arbejdsudbud: om det skyldes skiftende konjunkturer, ændringer i befolkningens størrelse, den aldersmæssige sammensætning, arbejdsmarkedsreformer og/eller ændrede institutioner. Betydningen af befolkningens størrelse kan delvis afhjælpes ved at benytte andele af den analyserede befolkningsgruppe i stedet for de absolutte tal. Effekten af arbejdsmarkedsreformer bør derfor primært måles på, om andelen af en betragtet befolkningsgruppes arbejdsmarkedstilknytning (erhvervs- og beskæftigelsesfrekvens) har ændret sig.

Arbejdsmarkedsdata fra Danmarks Statistik (DS) er det mest omfattende statistiske materiale, som står til rådighed. Men også her gælder det, at der er forskellige opgørelsesprincipper. Hertil kommer såkaldte 'brud' i de enkelte talserier, der i sig selv vanskeliggør en sammenhængende statistisk analyse omfattende en længere periode. Den 'officielle' arbejdsmarkedsstatistik blev senest omlagt fra 2008, hvilket giver en tidsserie på knap 15 år, som jeg helt overvejende vil benytte mig af. Dette udgangsår er dog ikke ganske uproblematisk, idet 2008 var et atypisk år med udpræget højkonjunktur. Noget tilsvarende kan måske siges om slutåret 2023, hvilket på den anden side kan bidrage til at belyse betydningen af efterspørgslen efter arbejdskraft for beskæftigelse og dermed arbejdsløshed.

Jeg vil afslutte dette indledende statistiske afsnit med at give et helt overordnet billede af udviklingen på det danske arbejdsmarked fra 2008 til 2023 ved brug af tal fra AKU-statistikken, der som nævnt (desværre) kun kan føres tilbage til 2008 på en konsistent basis.[4]

[4] Se mere detaljerede oplysninger vedr. AKU-statistikken:
https://www.dst.dk/da/Statistik/dokumentation/statistikdokumentation/arbejdskraftundersoegelsen--aku-

Tabel 1. Arbejdsmarkedsforhold, 2008-2023

AKU-statistik

Aldersgrupper: 15-74 år	2008	2013	2018	2023
1000 personer				
Beskæftigede	2809	2638	2833	3004
AKU-ledige	107	210	153	162
Uden for arbejdsstyrken	1177	1378	1369	1225
i alt	4093	4226	4355	4391
Procent				
Beskæftigelsesfrekvens	68,6	62,4	65,1	68,4
Ledighed, pct.	3,67	7,37	5,12	5,11
Erhvervsfrekvens[5]	71,2	67,4	68,6	72,1
Ginikoefficient	27,9	27,9	29,6	30,6

Noter:

1. AKU-statistikken er baseret på interviews og omfatter i denne opgørelse personer i alderen 15-74 år.
2. Gini-koefficienten er et beregnet tal mellem 0 og 100, jo højere tal desto mere ulighed i disponibel indkomst opgjort på husstand. (Det kan tolkes som et groft mål for, hvor stor en andel af den samlede disponible husstandsindkomst, der skal flyttes fra den halvdel af husstandene, der tjener mest til den halvdel, der tjener mindst, for at gøre indkomsten imellem alle befolkningsgrupper helt lige. I 202 er det således godt 30 pct. af den samlede disponible indkomst, der skulle omfordeles for at skabe en lige indkomstfordeling.

Kilde: Danmarks Statistik, Arbejdsmarkedsstatistik: AKU110A og IFOR41

Det ses af tabellen, at antallet af beskæftigede personer er steget med ca. 200.000 personer igennem de viste 15 år. Et tal der dog må vurderes i lyset af udviklingen i befolkningens størrelse, hvilket hhv. beskæftigelses- og erhvervsfrekvenserne belyser. Her ses det, at for perioden 2008-2023, hvor Danmarks Statistik angiver, at tallene er sammenlignelige, er erhvervs- og beskæftigelsesfrekvensen i både 2013 og 2018 lavere end i 2008. Dette resultat er mildt sagt overraskende, da det om noget har været skiftende regeringers erklærede mål gennem striben af pensions-, arbejdsmarkeds-, social- og skattereformer at øge andelen af befolkningens tilknytning til arbejdsmarkedet.

Som en konsekvens af arbejdsmarkeds- og skattepolitikken er der derimod igennem hele perioden sket en stigning i indkomstuligheden, der har flyttet Danmark fra at være et af de mest lige samfund i OECD til nu at tilhøre (den nedre del af) mellemgruppen af OECD-lande.

[5] Bemærk, at 'erhvervsfrekvensen' beregnes som (summen af antal beskæftigede + antal ledige) divideret med det samlede antal personer i de omhandlede aldersgrupper, 15-74 år.

Det analytiske grundlag for arbejdsmarkedsreformerne: arbejdsudbud skaber sin egen efterspørgsel - i DREAM-modellen

Når jeg indledningsvist ønskede at fremlægge de helt overordnede tal for udviklingen i arbejdsudbud og beskæftigelse, skyldtes det, at udviklingen i hhv. beskæftigelses- og erhvervsfrekvens for perioden 2008-2023 er overraskende – og efterlader umiddelbart det ubesvarede spørgsmål: 'hvilken effekt har arbejdsmarkedsreformerne haft på beskæftigelsen på trods af de omfattende reformer, der har ændret indkomstfordelingen markant?

Dette ubesvarede spørgsmål udfordrer således den simple regneregel, der benyttes i Finansministeriet (FM) ved beregning af udbudspolitikkens konsekvenser dvs. 'råderummet i dansk økonomi'. Ved beregningen af råderummet – opgjort som en hhv. positiv/negativ saldo på den offentlige sektors budget - antages det a priori, at **udbuddet af arbejdskraft og dermed beskæftigelsen har en substitutionselasticitet på 0,1** med hensyn til forskellen mellem real lønindkomst og arbejdsmarkedsrelaterede overførselsindkomster, begge opgjort efter skat[6]. Denne hypotetiske elasticitet benyttes efterfølgende i finansministeriets regnemodel til opgørelsen af ændringer i råderummet (dvs. den strukturelle budgetsaldo) for den økonomiske politik.

Arbejdsmarkedet i finansministeriets regnemodel er (ligeledes a priori) konstrueret således, at et øget udbud omsættes til øget beskæftigelse i forholdet 1:1. Det sker under henvisning til moderne makroøkonomisk teori, der er helt dominerende blandt såkaldt mainstream modeløkonomer Antagelsen er her, at i en velfungerende markedsøkonomi vil: *udbud [på arbejdsmarkedet] skabe sin egen efterspørgsel*[7]. Den makromodel, der benyttes i finansministeriet, har således denne postulerede egenskab, at udbud af arbejdskraft altid efterfølgende omsættes til en tilsvarende stigning i beskæftigelsen.[8]

---

[6] Substitutionselasticiteten og indkomstfølsomheden forudsættes ens for alle beskæftigede, det vil sige uafhængig af køn, arbejdstid og indkomstniveau.
Ministerierne anvender en substitutionselasticitet på 0,1 og en indkomstfølsomhed på -0,05 på tværs af alle beskæftigede, Finansministeriet, Regneprincipper, 2024, p.6

[7] Denne antagelse er kanoniseret blandt neoklassiske makroøkonomer, jfr. f.eks. tidligere overvismand Hans Jørgen Whitta-Jacobsen udsagn i 2011 (midt under finanskrisen) gengivet nedenfor, og som jeg spurgte ham om for blot et år siden, om han fortrød. Det gjorde han ikke.

[8] Finansministeriets regnemodel er en tilpasset version af ADAM-modellen, oprindeligt udviklet i Danmarks Statistik på initiativ af professor Ellen Andersen. Den blev overtaget og efterfølgende stærkt ændret af finansministeriets økonomer, i en grad så Ellen Andersen havde svært ved at genkende den, idet hun udtalte i 2017: "Modellen siger det, der bliver puttet ind i den. Svaret er ikke noget, der kommer ud af den, det er noget, der kommer ind i den. Og det er ikke inde i ADAM. Det kommer fra Finansministeriet." (Professor Ellen Andersen, 27. februar 2017 interview i anledning af 80 årsdagen med Kristine Dons Christensen, Zetland), Regnemodellernes moder advarer mod at stole for meget på den model, hun selv har skabt (zetland.dk).

Men utilfredsheden i finansministeriet med ADAMs grundlæggende egenskaber tog til i et omfang, så det besluttede at opgive samarbejdet med Danmarks Statistik om modeludvikling. Der var stigende uenighed om tilpasningshastigheden på arbejdsmarkedet, så det blev besluttet at opbygge en helt ny regnemodel, MAKRO. Konsulentfirmaet DREAM blev engageret til at forstå dette arbejde under ledelse af en styregruppe bestående af repræsentanter fra offentlige institutioner, herunder naturligvis finansministeriet og udvalgte modeløkonomer fra de højere læreanstalter, se bl.a. DREAM, 2019 og Jespersen, 2017. Det i stigende grad blev vanskeligt at få de statistiske estimationer, som Danmarks

Men denne antagelse matcher ikke den faktiske udvikling, hverken totalt set som beskrevet ovenfor i tabel 1 eller for arbejdsmarkedets kernegrupper (aldersmæssigt: 35-55 år) igennem de seneste 15 år, jfr. tabel 2 nedenfor. Derimod synes reduktion af de arbejdsmarkedsrelaterede sociale ydelser og mindsket progression i skatteskalaen at have den tilsigtede – omend ikke 1:1 – effekt med hensyn til at øge indkomstuligheden, jfr. Velfærdskommissionens redegørelse, kap. 4.

Et ændret udbud af arbejdskraft er *ikke* ensbetydende med en tilsvarende ændring af beskæftigelse

Forskellen mellem arbejdsudbud og beskæftigelse udgøres af det ofte ganske betydelige og varierende antal personer, der i kortere eller længere perioder – i ordets egentligste forstand – er arbejdsløse. Et tal, der svinger både med konjunkturerne, med den økonomiske politik og med demografien; men som helt åbenbart ikke forsvinder af sig selv.

Går vi tilbage til begyndelsen af 1990erne var der også forskel mellem den registerbaserede og den AKU-opgjorte arbejdsløshed. Men dengang lå det registerbaserede tal højere en AKU-statistikken, hvilket kunne tolkes som, at der må have været en del modtagere af dagpenge og kontanthjælp, der ikke ifølge interviewundersøgelserne umiddelbart stod til rådighed for arbejdsmarkedet. For så vidt ikke et overraskende resultat i og med at arbejdsløsheden lå på ca. 10 pct. - svarende til ca. 300.000 personer, der reelt stod uden arbejde. Det tal blev op gennem 1990erne reduceret ganske betydeligt gennem en målrettet beskæftigelsespolitik (efterspørgsel) og en målrettet orlovs- og efteruddannelsespolitik kombineret med en reduktion af ledighedsydelsernes størrelse og varighed, hvorved der overgangsvist kom balance mellem den register- og den interviewbaserede arbejdsløshedsstatistik. Efter regeringsskiftet i 2001 og med afsæt i Velfærdskommissionens (2003-05) anbefalinger blev der som nævnt gennemført en stribe omfattende arbejdsmarkedsreformer med det primære sigte at øge udbuddet af arbejdskraft. De væsentligste instrumenter hertil var en forøgelse af pensionsalderen, kraftig beskæring af efterlønnen, forøgelse af forskellen mellem og omfanget af ledighedsydelser i forhold til den disponible indkomst, som også en stribe indkomstskattelettelser bidrog til.

Statistik stadig var hovedleverandør af til at matche de principper, som finansministeriet ønskede, at dets regnemodel burde baseres på, bl.a. med hvilken hast arbejdsmarkedsreformer slog igennem på beskæftigelsen og dermed råderummet i den økonomiske politik.

Hvad viser virkeligheden….

Benyttes AKU-statistikken er udviklingen igennem de seneste 15 år tankevækkende. Den positive effekt på beskæftigelsen af den stribe af udbudsforøgende 'reformer', som kommissionsøkonomerne anbefalede, og som politikerne derfor fik 'lovning på', hvis de reducerede de sociale ydelser og mindskede progressionen i skatteskalaen, er svær at genfinde i statistikken, navnlig for de 35-54 årige – der sædvanligvis betegnes som arbejdsmarkedets 'kernetropper', se tabel 2.

Jeg har på grund af de ovenfor beskrevne ændringer i statistikken valgt at benytte den definitorisk sammenlignelige periode begyndelsen i 2008 til slutningen i 2022 (der begge repræsenterer to konjunkturtoppe)[9], jfr. tabel 2. Her kan det konstateres, at beskæftigelsesfrekvensen set under ét *ikke er steget* – ja, den er faktisk ½ pct. point lavere i 2022 end i 2008. Denne udvikling dækker ydermere over en skævtrækning af udviklingen, når den opdeles på aldersgrupper, idet arbejdsmarkedstilknytningen for alle grupper op til en alder af 45 år er *faldet*. For aldersgruppen fra 45 til 54 år er tilknytningen – uanset de mange reformer - stort set uændret. Det er først for grupperne 55 år og opefter at der ses en klar beskæftigelses- og udbudseffekt, hvilket ikke er overraskende, når de betænkes, hvorledes førtidspension er blevet markant forringet, og aldersgrænsen for at opnå folkepension er blevet forøget med to år.

Hvad kan man slutte heraf? 1. udbuddet af arbejdskraft må (også) være bestemt af andre forhold end størrelsen og varigheden af de sociale ydelser og progressionen i skatteskalaen for personer i disse aldersgrupper, der som arbejdsmarkedsparat ikke har mulighed for at få en anden social ydelse. Man kunne ligefrem få den kætterske tanke, at forringelserne for de yngre aldersgrupper ligefrem har haft den modsatte effekt – at de har trukket personer ud af arbejdsstyrken i takt med, at retten til en social ydelse ophørte, og jobmulighederne begrænset. (Udfaldseffekten fra arbejdsstyrken var i hvert fald overraskende stor ca. 35.000 personer i 2013, da den reducerede dagpengeret fra 4 til 2 år var fuldt implementeret. En ret der kun havde omfattet personer i dagpengesystemet, altså personer med en betydelig tilknytning til arbejdsmarkedet, se Økonomi- og Indenrigsministeriet, 2019.

Ovenstående peger entydigt i retning af, at den hidtil i finansministeriet benyttede tommelfingerregel, at der uanset aldersgruppe er en arbejdsudbudselasticitet på 0,1 knyttet til forskellen mellem den disponible lønindkomst og arbejdsmarkedsydelser (efter skat) trænger til et virkelighedstjek. Og det samme kunne beregningsmodellens a priori-antagelse, at udbud af arbejdskraft skaber sin egen efterspørgsel efter arbejdskraft, også med fordel underkastes, se finansministeriet, 2024

Men det ville være uretfærdigt at skære alle neoklassiske økonomer over en kam. Det er således bemærkelsesværdigt, at Rockwool Fondens forskningsenhed - uanset at den

[9] Uanset, at allerede i 2019 var der en udbredt opfattelse af, at 'Lige nu er der bragende højkonjunktur i Danmark' Leif Bech Fallesen skrev 3. okt. 2019 i Politiken.

tidligere har været en stærk fortaler for de gennemførte arbejdsmarkedsreformer – i en efterfølgende evaluering har nuanceret sin tidligere konklusion, at nedskæring af de sociale ydelser har haft en positiv beskæftigelseseffekt, se forskningsrapporten *Hvordan udvikler beskæftigelsen sig i Danmark?, juni 2019*. Forskerne har her primært benyttet data fra *Styrelsen for Arbejdsmarked og Rekruttering*. Efter at have gransket disse tal må forskerne om end med en vis tøven konstatere, at netop for perioden fra 2008 til 2018 – altså fra en konjunkturtop til den næste konjunkturopgang, hvor arbejdsmarkedsreformerne har haft mulighed for at virke i op mod ti år - er antallet af (fuldtids)beskæftigede danskere i alderen 16-64 år *faldet* med ikke mindre end 150.000 personer. Rockwool Fonden formulerer resultatet således: 'det er særligt de 25-54 årige (kernegrupperne), der har mistet terræn i forhold til andre EU-lande'.[10]

Og hvordan er det gået for personer, der fik (top)skattelettelser? Her er der også modsatrettede effekter i spil, hhv. en substitutionseffekt mellem arbejde og fritid og en indkomsteffekt (jo større disponibel indkomst desto mindre incitament til at arbejde mere). Jens Bonke og Schultz-Nielsens undersøgelse fra 2014 giver en indikation af, hvilken af effekterne der dominerer. Her konstateres det, at der er en majoritet af danskere, der ønsker at reducere det ugentlige antal arbejdstimer, men her stiller den overenskomstmæssigt aftalte arbejdstid sig hindrende for reduktion af den ugentlige arbejdstid. Et resultat der flugter med det overordnede billede af udviklingen i det gennemsnitlige antal årligt udførte arbejdstimer pr. person, gengivet i tabel 1.

Derimod er udbud og beskæftigelse steget for de ældste aldersgrupper (55+), hvor reformerne har været langt mere omfattende, idet den aldersbetingede ret til en permanent indkomsterstattende ydelse bortfaldt for de aldersgrupper, der blev omfattet af forhøjelsen af pensionsalderen, beskæring af efterløn og stærkt reduceret mulighed for at få tilkendt permanent førtidspension. Disse reformer har samlet set indebåret en betydelig forøgelse af pensionsalderen, hvilket i kombination med et generel løft i den aldersrelaterede sundhedstilstand har øget både erhvervs- og beskæftigelsesfrekvenserne for aldersgrupper over 55 år. Bemærkes bør det, at disse frekvenser rent faktisk begyndte at stige, *før* disse reformer havde fuld effekt[11]. En udvikling der ligeledes må tolkes som, at adskillige andre forhold spiller en væsentlig rolle for arbejdsmarkedsdeltagelsen også i de ældste

[10] Her bør det også nævnes, at konsulentgruppen DREAM, der arbejder tæt sammen med finansministeriet, i et forsøg på at bekræfte deres hidtil mekaniske brug af regnereglen omfattende en udbudselasticitet på 0,1 og fortsat klare anbefaling af arbejdsmarkedsreformerne heller ikke kan påvise en positiv effekt for de nævnte kernegrupper på arbejdsmarkedet. Og det på trods af, at forskerne gør sig ihærdige anstrengelser for at påvise en sådan positiv effekt: idet 2018 sammenlignes med 2000 (der ikke var udpræget højkonjunktur), og der ydermere 1. justeres kraftigt for et 'databrud' i 2003; 2. der korrigeres for øget antal studerende; i hvilket omfang det reducerer arbejdsudbuddet er ikke klart belyst; 3. der korrigeres for en ændret befolkningssammensætning flere 1. og 2. generations indvandrere – også her er det uklart om den tilsyneladende lavere beskæftigelsesfrekvens er et udslag af lavere udbud eller snarere udtryk for, at disse grupper står bagest i arbejdsløshedskøen (efterspørgselseffekt).
Men på trods af disse tre 'korrektioner', der alle har til formål at øge beskæftigelsesfrekvensen i 2018 – lykkes det IKKE for DREAM-gruppen at påvise en forøget beskæftigelsesfrekvens for kernegrupperne, DREAM, 2019

[11] Pensionsalderen blev hævet fra 65 år til 67 i perioden fra 2019 til 2022. Hvilket - blot nævnt som et kuriosum - også var pensionsalderen for opnåelse af folkepension fra den blev indført i 1957 og frem til 2002.

aldersgrupper end blot de sociale ydelser. Disse forhold har også afspejlet sig i den betydelige stigning fra 2020 i antallet tildelte senior- og Arne-pensioner. Førstnævnte er en helbredsbetinget ydelse, der skal baseres på en lægelig vurdering, mens Arne-pensionen er rettighedsbetinget (bestemt af antal år på arbejdsmarkedet). Det er dog klart den helbredsbetingede pensionsret, der har været dominerende af de to (førtids)pensionsmuligheder. Et forhold, der måske også kan forklares ved at ydelsen for seniorpension er noget højere. Dette spørgsmål afventer en nærmere undersøgelse.[12]

Endelig bør udviklingen i de yngre aldersgrupper også vurderes i lyset af ovenstående. Det forhold af de ældste aldersgrupper nærmest bliver tvunget til at blive længere på arbejdsmarkedet i takt med, at pensionsalderen hæves, må indebære en risiko for, at der som en konsekvens heraf vil være færre ledige stillinger, der kan besættes af de unge, der er klar til at træde ind på arbejdsmarkedet. Dette omtales i dele af litteraturen, som en 'omvendt' gøgeunge-effekt, hvor de ældre skubber de yngre årgange ud af arbejdsmarkedet. Det skal dog også bemærkes, at den mindskede arbejdsmarkedstilknytning for de yngre årgange også skyldes en øget tilgang til ungdoms- og videregående uddannelser.

Afslutning

Som det er fremgået af ovenstående gennemgang af udviklingen på det danske arbejdsmarked igennem de seneste årtier, er den kendetegnet ved en stribe ganske omfattende reformer, der over en bred kam kan karakteriseres som en mindskelse og tidsmæssig afkortning af de arbejdsmarkedsrelaterede ydelser, en reduktion i progressionen i indkomstskatten og en forhøjelse af pensionsalderen. Uanset disse ændringer kan der ikke i den selvsamme periode påvises en signifikant stigning i arbejdsmarkedstilknytningen total set; men derimod en markant stigning i indkomstuligheden målt ved Ginikoefficienten.

Ses der mere detaljeret på udviklingen i de enkelte (tiårs) aldersgrupper, fremtræder der dog en afgørende forskel. Arbejdsmarkedets kernetropper i alderen fra 35-55 år har siden 2008 fortsat en høj, men reduceret(!) tilknytning til arbejdsmarkedet. Aldersgrupperne over 55 har derimod opnået en markant forøget erhvervs- og beskæftigelsesfrekvens, der delvis kan henføres til forøgelsen af pensionsalder, et (delvist) bortfald af efterløn og stærkt beskåret adgang førtidspensionering. Et billede som muligheden for opnåelse af seniorpension og Arne-pension ikke har ændret.

[12] Efterfølgende har det også vist sig, at finansministeriet overvurderede det antal personer, der ville benytte sig af Arne-pensionen – hvilket blot understreger betydningen af, at der er jobs til rådighed, som en væsentlig determinant ved valget med arbejdsindkomst og arbejdsmarkedsydelse, selv af permanent karakter. https://fm.dk/media/18137/ny-ret-til-tidlig-pension-vaerdig-tilbagetraekning-for-alle.pdf?fbclid=IwAR14WkDi0KrReUzIIezjtiiL6Xe9LJ-SyyEEP4MtoFhRKSQAxBwNhLCN5FA

Ovenstående efterlader et flimrende billede af hvilke faktorer, der er dominerende med hensyn til udviklingen på arbejdsmarkedet, herunder hvilken rolle de økonomiske incitamenter egentlig spiller. Men konstateres kan det, at der ikke synes at være nogen simpel sammenhæng, endsige tommelfingerregel, fra ændringer i social- og skattepolitikken til udbuddet af arbejdskraft, eller fra udbuddet af arbejdskraft (registreret eller skjult) til beskæftigelsen. Her spiller efterspørgslen efter arbejdskraft fortsat en betydelig rolle.

Litteratur:

Andersen, T.M. : *Velfærdskommissionen – formål, resultater og erfaringer,* Samfundsøkonomen, 2/2021, s. 25-35

Bonke, J. og M.L. Schultz-Nielsen: *Do Preferences Impact Behavior and Wellbeing? A Panel Study of Preferred and Actual Working Time 2001-2008/09*, *IZA Discussion Paper No. 8356*, 34 Pages Posted: 2 Aug 2014

Danmarks Statistik: Danmarks Statistiks forskellige ledighedsbegreber, 2014, https://www.dst.dk/Site/Dst/SingleFiles/GetArchiveFile.aspx?fi=arbe&fo=ledighedsbegreber--pdf&ext={2}

DREAM-gruppen: *Ændringer i erhvervsdeltagelsen siden årtusindeskiftet*, arbejdspapir, 2019:4

DREAM-gruppen: *Effekterne af permanent øget arbejdsudbud*, dec. 2023 https://dreamgruppen.dk/Media/638448965067506331/arbejdsudbudsstoed.pdf

Finansministeriet, 2024 Regneprincipper personskatter, https://fm.dk/media/mumhodwc/regneprincipper-paa-personskatteomraadet.pdf

Greve, B. m.fl., *Arbejdsmarkedet og arbejdsmarkedspolitik - en kritisk analyse*, København: Samfundsfagslitteratur, 1996

Grønnegaard Christensen, J. m.fl.: *De store Kommissioner – vise mænd, smagsdommere eller nyttige idioter?* Syddansk Universitets Forlag, 2009

Jespersen, J.: *Økonomien, journalistikken og den udeblevne kritik* i R. Buch & M. Verner (red.), *Krisen i økonomi & Journalistik,* Forlaget AJOUR, 2014

Jespersen, J.: *Vækstøkonomi på Vildspor,* Essaysamling, Forlaget Jensen&Dalgaard, 2019

Jespersen, J.: *Arbejdsmarkedsreformer på godt og ondt*, Særnummer: Velfærdsstatens udfordringer, forandringer og konsekvenser, Samfundsøkonomen, 2023/4, s. 16-26

Regeringen: *Ny ret til tidlig pension*, august 2020

Rockwoolfonden: *Hvordan udvikler beskæftigelsen sig i Danmark?, juni 2019*

Skatteministeriet: *Skattereform, beskæftigelse og velfærd*, 2012

Velfærdskommissionen: *Fremtidens velfærd kommer ikke af sig selv*, Analyserapport, 2004

Velfærdskommissionen: *Fremtidens Velfærd*, *slutrapport*, 2006, file:///C:/Users/jesperj/Downloads/Velf%C3%A6rdskommissionen+%E2%80%93+form%C3%A5l,+resultater+og+erfaringer.pdf

Økonomi- og Indenrigsministeriet: *Dagpengereformen 2010: Hvordan gik det?* Økonomisk Analyse, 25-02-2019

Tabel 2: Beskæftigelses- og erhvervsfrekvens og arbejdsløshed, 2008-2022

		2008	2013	2017	2022
Beskæftigelsesfrekvens	**Alder i alt**	68,6	62,4	64,2	68,2
	15-24 år	62,7	49,6	53,0	56,3
	25-34 år	84,6	76,1	75,2	80,5
	35-44 år	88,5	84,4	83,9	85,8
	45-54 år	87,3	82,6	84,7	87,6
	55-64 år	56	58,8	68,2	72,9
	65-74 år	11,2	12,9	13,8	17,1
AKU-ledighedsprocent	**Alder i alt**	3,7	7,4	5,8	4,5
	15-24 år	9,5	14,8	12,4	10,5
	25-34 år	3,7	8,9	7,6	5,4
	35-44 år	2,6	5,8	4,6	3,2
	45-54 år	2	5,3	3,6	2,3
	55-64 år	2,5	5,5	3,7	2,9
	65-74 år	..	..	..	2,8
Erhvervsfrekvens	**Alder i alt**	71,3	67,4	68,2	71,4
	15-24 år	69,2	58,2	60,4	63
	25-34 år	87,8	83,6	81,4	85,1
	35-44 år	90,8	89,6	87,9	88,6
	45-54 år	89,1	87,3	87,8	89,7
	55-64 år	57,5	62,2	70,8	75,1
	65-74 år	11,3	12,9	13,9	17,6

Kilde: Danmarks Statistik, AKU111A

Økonomers køn og løn

En statistisk sammenligning af kvindelige og mandlige cand. polit'ers lønfordelinger

Nadja Eifler, Rockwool Fondens Interventionsenhed
Henrik Hansen, Økonomisk Institut på Københavns Universitet

Resumé

Vi analyserer kønsforskelle i løn på tværs af hele lønfordelingen blandt nyuddannede økonomer fra Københavns Universitet, som alle arbejder fuld tid. Vi finder, selv her, betydelige lønforskelle. Forskellene kan ikke forklares af observerbare karakteristika, så som karakterer, erhvervserfaring i studietiden, studietid, sektorvalg, ansættelse i højtlønnede brancher, eller børn. I den offentlige sektor er den uforklarede lønforskel mellem mænd og kvinder på næsten 7.000 kr. ved lønfordelingens 10. percentil, mens den er 21.000 kr. ved lønfordelingens 90. percentil. Forskellene er dog kun marginalt siknifikante. I den private sektor er den uforklarede lønforskel statistisk signifikant på tværs af hele fordelingen, og på 17.000 kr. ved 10. percentil, mens den er 86.000 kr. ved 90. percentil. Forskelle i jobfunktioner eller familierelaterede forskelle kan muligvis forklare en del af kønseffekten, men sådanne forskelle kan ikke forklare, hvorfor en gruppe unge mænd og kvinder, der har taget de samme lange uddannelsesvalg, pludselig begynder at divergere i deres valg umiddelbart efter de forlader deres studie.

1. Indledning

Mænd og kvinder i Danmark får forskellig løn. De politiske og videnskabelige debatter fokuserer på, i hvilket omfang lønforskellene kan tilskrives de valg, mænd og kvinder træffer i deres studieliv, familieliv og arbejdsliv. Hvis lønforskellene skyldes forskellige valg, er der ikke et egentligt ligelønsproblem, da ligeløn principielt betyder lige løn for lige arbejde. Lønforskelle, der derimod ikke er et resultat af frivillige valg, udgør et problem, idet de repræsenterer en systematisk forskelsbehandling på arbejdsmarkedet, som kan tilskrives individernes køn. Derfor opdeler økonomer traditionelt lønforskellen i to; en forklaret og en uforklaret del (se Blau & Kahn, 2017; Kunze, 2017 eller Larsen mfl., 2020). I denne artikel benævner vi de to dele hhv. "karakteristika" og "kønseffekten". Karakteristika henføres til forskelle i uddannelse, jobfunktioner, erhvervserfaring, sektor- og branchevalg samt, i stigende omfang, familiens arbejdsdeling. Disse observerbare forskelle, som kan påvirke individuelle lønninger, tager man højde for i økonometriske analyser. Den forskel der måtte være tilbage i gennemsnitslønningerne for mænd og kvinder efter der er taget højde for de observerbare forskelle, kønseffekten, har traditionelt været tolket som et udtryk for en systematisk forskelsbehandling på arbejdsmarkedet.

Nyere danske studier af lønforskellen mellem mænd og kvinder bruger registerdata fra Danmarks Statistik, og de baserer sig typisk på mange personer, fx 1.848.200 lønmodtagere i Larsen et al. (2020). I Eifler & Hansen (2024) går vi den modsatte vej, idet vi alene ser på dimittender fra økonomiuddannelsen på Københavns Universitet (cand. polit'er). Vi estimerer karakteristika og kønseffekter i gennemsnitslønningerne for cand.polit'er, der dimitterede i perioden 1994-2013. Vi opdeler denne gruppe efter deres ansættelse i enten den offentlige eller den private sektor, og vi tager via regressioner højde for forskelle i deres evner og kundskaber i det omfang det afspejles i deres karaktergennemsnit på polit-uddannelsen, og deres opnåede erhvervserfaring i studietiden, og vi medtager oplysninger om, hvorvidt de har børn. Vi sammenligner dermed lønninger for en lille gruppe mænd og kvinder, som har meget ens karakteristika. Sammenligningen viser, at forskellen i gennemsnitslønningerne to år efter dimissionen, som kan henvises til observerbare forskelle i evner og personlige valg, er lille. Broderparten i lønforskellene er således kønseffekten. Denne forskel er betydelig, både i den offentlige sektor (ca. 16.000 kr. om året) og i den private sektor (ca. 40.000 kr. om året).

Nærværende artikel bygger videre på Eifler & Hansen (2024), idet vi udvider samplet, og samtidig foretager en langt grundigere analyse, i det vi sammenligner løn*fordelingerne* for kvinder og mænd. Analysen viser, at lønforskellene er markant stigende over lønfordelingerne. For polit'er ansat i den offentlige sektor stiger lønforskellen fra godt 6.000 kr. i den lave ende af fordelingerne til ca. 30.000 i den høje ende. Der er dermed tale om en stigende forskel i både absolut og relativ forstand. Samtidig kan vi ikke afvise, at hele lønforskellen i den offentlige sektor er en kønseffekt. I den private sektor finder vi ligeledes stigende forskelle over lønfordelingerne, fra ca. 25.000 kr. i den nedre ende til mere end 100.000 kr. i den høje del af fordelingen. Til forskel fra resultaterne i Eifler & Hansen (2024) kan vi med det nye udvidede sample ikke længere afvise, at hele lønforskellen i den private sektor er en kønseffekt, idet effekten af forskelle i karakteristika kun er marginalt signifikant på et 10% signifikansniveau.

Artiklen er struktureret som følger. I næste afsnit viser vi, hvordan vi identificerer og estimerer effekterne af forskelle i karakteristika samt kønseffekterne på tværs af lønfordelingerne. I afsnit 3 beskriver vi, hvordan vi har udvalgt vores sample af cand. polit'er fra registrene i Danmarks Statistik. Vi har samlet alle empiriske resultater i afsnit 4, og vi afslutter artiklen med et par kommentarer i afsnit 5.

2. Identifikation og estimation af kønseffekter

2.1. Lønfunktioner og lønfordelinger

Vi antager, at lønnen (W_i) for hver cand. polit er et resultat af observerbare og uobserverbare komponenter (X_i, ε_i), specificeret af lønfunktioner, $g_k(\cdot,\cdot)$, hvor indekset k, angiver personens køn med 0 for kvinder og 1 for mænd:

$$W_{ik} = g_k(X_{ik}, \varepsilon_{ik}), \quad i = 1, \ldots N_k, \; k = 0,1. \tag{1}$$

I langt de fleste analyser af individuelle lønninger, antages den betingede middelværdi af lønfunktionen at være log-lineær, og de oftest inkluderede observerbare komponenter er uddannelse og erfaring (se fx, Blau & Kahn, 2017, eller Larsen mfl., 2020). I denne analyse vælger vi en lineær approksimation, idet vi har individer med samme uddannelse og meget ens erfaring. Samtidig ser vi på hele lønfordelingen og ikke alene den betingede middelværdi.

Den væsentlige antagelse for analysen er, at vi betragter løn (W), karakteristika (X) og køn (K) som stokastiske variable med en simultan fordeling, hvorfra vi kan udlede marginale og betingede fordelinger. De tre centrale fordelinger er hhv. de to, hvor vi har observationer: (i) Lønfordelingen for kvinder, $F_{00} = F(W_0 \mid K = 0)$, (ii) lønfordelingen for mænd, $F_{11} = F(W_1 \mid K = 1)$, samt (iii) den uobserverbare kontrafaktiske lønfordeling, vi ville have observeret for kvinder, hvis de observerede karakteristika havde været som for mænd, $F_{01} = F(W_0 \mid K = 1)$.

Den kontrafaktiske lønfordeling kan estimeres givet to antagelser. Den første betegnes enten som ignorabilitet (Rosenbaum & Rubin, 1983), selektion på observerbare karakteristika (Heckman & Robb, 1985) eller betinget uafhængighed (Lechner, 1999). Grundlæggende betyder antagelsen, at de uobserverede komponenter i lønfunktionerne (ε) givet de observerbare karakteristika (X) er uafhængige af køn (K). Den anden antagelse er overlappende støtte, hvilket betyder at der ikke er nogen værdier af karakteristika (X), som kun observeres for enten kvinder eller mænd. De to antagelser samles ofte i en fælles antagelse om *streng ignorabilitet* (Rosenbaum & Rubin, 1983).

Antagelsen er tilstrækkelig for estimation af de tre fordelinger, F_{00}, F_{11} og F_{01}. Specifikt, viser Firpo (2007), at vi kan bruge tre vægtfunktioner givet ved:

$$\omega_{11}(k, p(x)) = k \,/\, p,$$

$$\omega_{00}(k, p(x)) = (1-k) \,/\, (1-p),$$

$$\omega_{01}(k, p(x)) = \big((1-k) \,/\, (1-p(x))\big) \times \big(p(x) \,/\, p\big),$$

hvor (p) er sandsynligheden for, at en person er en mand, mens (1 - p) er sandsynligheden for, at vedkommende er en kvinde, og $p(x) = \mathrm{P}(K = 0 \mid X = x)$ er sandsynligheden for, at personen *kunne være* en kvinde, givet de observerbare karakteristika (dvs. propensity scoren). De to første vægtfunktioner er velkendte, idet de bruges til at danne de marginale lønfordelinger for hhv. mænd og kvinder, mens den tredje estimerer den kontrafaktiske fordeling under antagelsen om streng ignorabilitet. De tre fordelinger er givet ved:

$$F_{ab}(w) = \mathrm{E}[\omega_{ab}(k, p(x)) \cdot \mathbf{1}_{[W \le w]}], \quad a, b = 0,1. \tag{2}$$

I den empiriske analyse estimerer og sammenligner vi percentiler af de tre fordelinger. Lad derfor forskellen i percentil 'v' mellem fordelingen for mænd og fordelingen for kvinder være givet ved:

$$\Delta^{\nu} = \nu(F_{11}) - \nu(F_{00}) \tag{3}$$

Analogt til en Oaxaca-Blinder-dekomponering (Blinder, 1973; Oaxaca, 1973) af forskelle i gennemsnitslønningerne, opdeles forskellen i percentilerne i en kønseffekt og karakteristika ved brug af den kontrafaktiske fordeling

$$\Delta^{\nu} = [\nu(F_{11}) - \nu(F_{01})] + [\nu(F_{01}) - \nu(F_{00})] = \Delta_K^{\nu} + \Delta_X^{\nu} \tag{4}$$

Firpo (2007) viser, at disse funktionaler af de tre fordelinger er identificeret givet streng ignorabilitet og en antagelse om entydige percentiler. Δ^{ν}_X er forskellen i percentilerne, som kan henføres til forskelle i de observerede karakteristika, mens kønseffekten (Δ^{ν}_K) er forskellen i percentilerne, som skyldes forskellen i de to køns lønfordelinger. Dette forbinder lønfunktionerne med opdelingen af percentilerne, fordi ens lønfunktioner for de to køn medfører, at kønseffekten er nul.

2.2. Identifikation og estimation af kønseffekten med betydningsfunktioner

Firpo m.fl. (2009) foreslår en regressionsbaseret estimation af forskelle i visse funktionaler af fordelinger ved brug af re-centrerede betydningsfunktioner (engelsk: recentered influence functions, RIF).[1] RIF'erne for percentiler kan gives som:

$$H_{ab}(W,\tau) = q_{ab}(\tau) + (\tau - \mathbf{1}_{[W \le q_{ab}(\tau)]}) / f_{ab}(q_{ab}(\tau)), \quad a,b = 0,1, \tag{5}$$

hvor $q_{ab}(\tau) = \inf\{w \mid F_{ab}(w) \ge \tau\}$ er den τ-te percentil i fordelingen F_{ab}, mens f_{ab} er tæthedsfunktionen og $\mathbf{1}_{[.]}$ er indikatorfunktionen. For hver percentil er $H_{ab}(W,\tau)$ en stokastisk variabel, og RIF-regressionerne er de betingede middelværdier givet ved:

$$\mathrm{E}(H_{ab}(W,\tau) \mid X = x) \equiv m_{ab}(x,\tau), \quad a,b = 0,1 \tag{6}$$

RIF-regressionerne kan bruges til at estimere kønseffekterne i lønfordelingerne, fordi betydningsfunktionerne i sig selv har middelværdi 0, så $\mathrm{E}[H_{ab}(W,\tau)] = q_{ab}(\tau)$, og tårnegenskaben giver at:

$$\Delta_K^{q_\tau} = \mathrm{E}[m_{11}(X,\tau) \mid K = 1] - \mathrm{E}[m_{01}(X,\tau) \mid K = 1],$$
$$\Delta_X^{q_\tau} = \mathrm{E}[m_{01}(X,\tau) \mid K = 1] - \mathrm{E}[m_{00}(X,\tau) \mid K = 0].$$

Hvis de betingede middelværdier approksimeres med lineære funktioner

$$m_{ab}(x,\tau) = x'\beta_{ab}(\tau), \quad a,b = 0,1,$$

kan parametrene estimeres med almindelig regression:

$$\beta_{00}(\tau) = \left(\mathrm{E}[X_0 X_0' \mid K = 0]\right)^{-1} \mathrm{E}[X_0 H_{00}(W,\tau) \mid K = 0]$$

[1] Betydningsfunktioner (Influence functions) defineres første gang i Hampel (1974). Funktionerne kvantificerer, hvordan en given statistik ændres, når en lille mængde datamasse tilføjes til et bestemt punkt i fordelingen, den givne statistik er baseret på. Man kan også tænke på betydningsfunktionen som en approksimation af betydningen af en observation på en statistik beregnet ud fra et datasæt. Alle betydningsfunktioner har middelværdi nul. Recentreringen i Firpo m.fl. (2009) giver dermed betydningsfunktionen samme middelværdi som den oprindelige statistik.

$$\beta_{01}(\tau) = \left(\mathrm{E}[X_0 X_0' \mid K=1]\right)^{-1} \mathrm{E}[X_0 H_{01}(W,\tau) \mid K=1]$$
$$\beta_{11}(\tau) = \left(\mathrm{E}[X_1 X_1' \mid K=1]\right)^{-1} \mathrm{E}[X_1 H_{11}(W,\tau) \mid K=1]$$

Som beskrevet i Firpo m.fl. (2018) leder dette til en opdeling af forskellen i percentilerne, som er meget lig en Oaxaca-Blinder-dekomponering af forskellen i gennemsnittene i en lineær regressionsmodel:

$$\Delta_K^{q_\tau} = \mathrm{E}[X_1 \mid K=1]'(\beta_{11}(\tau) - \beta_{01}(\tau)) + \mathrm{E}[X_1 - X_0 \mid K=1]'\beta_{01}(\tau) \quad (7)$$

$$\Delta_X^{q_\tau} = (\mathrm{E}[X_0 \mid K=1] - \mathrm{E}[X_0 \mid K=0])'\beta_{00}(\tau) + \mathrm{E}[X_0 \mid K=1]'(\beta_{01}(\tau) - \beta_{00}(\tau)) \quad (8)$$

I (7) er det første led kønseffekten, mens det andet led er et fejlled, som skyldes den potentielle forskel i de observerbare karakteristikas gennemsnit, når de vægtes med henholdsvis ω_{11} og ω_{01}. I (8) er første led effekten af forskelle i karakteristika, mens andet led er en approksimationsfejl, der fremkommer, hvis der er forskelle i fordelingsfunktionernes kurvatur.

Fordelen ved RIF-regressionsmetoden fremgår af (7) og (8). Med et givent løndatasæt er ligningerne to Oaxaca-Blinder-dekomponeringer af funktioner af tre vægtede datasæt. Kønseffekten i (7) er en sammenligning af de observerede lønninger for mænd med re-vægtede løndata for kvinder, mens karakteristika-effekten i (8) er en sammenligning af de re-vægtede lønninger med de observerede data for kvinderne.

3. Datagrundlag og -behandling

Danmarks Statistiks registerdata har oplysninger om 4.160 dimittender fra polit-studiet i årene 1994-2017. Af disse medtager vi kun dimittender, som har både bachelor- og kandidatuddannelsen fra polit-studiet, og vi udelader dimittender med meget korte og meget lange indskrivninger, så alle dimittender har været på studiet i mindst 3 år og højest 10 år. Blandt disse medtager vi alene dimittender, der er i fuldtidsbeskæftigelse i det andet år efter deres dimission. Denne udvælgelse reducerer populationen til 2.915 dimittender, 2.003 mænd og 912 kvinder. Afgrænsningen afspejler at vi ønsker at beskrive de kønsmæssige lønforskelle blandt unge fuldtidsansatte dimittender, man almindeligvis vil karakterisere som typiske cand. polit'er.

Tabel 1 viser de gennemsnitlige årlige lønninger målt i faste 2015-priser, samt en række karakteristika for de 2.915 dimittender fordelt på køn og sektor, hvor sektoren er givet af dimittendernes ansættelsessted i deres andet år efter dimission. Lønindkomsten er variablen LOENMV_13 fra Danmarks Statistiks register for personindkomst.

De velkendte lønforskelle mellem ansatte i den private og den offentlige sektor er markante. Samtidig ses betragtelige, og statistisk signifikante lønforskelle mellem mænd og kvinder i begge sektorer. Da 65% af mændene er ansat i den private sektor i forhold til 42% af kvinderne, kan en del af den gennemsnitlige lønforskel mellem mandlige og kvindelige dimittender på 59.000 kr. henføres til sektorvalg. Samtidig ses, at der inden

for sektorerne også er signifikante løngab: 18.000 kr. i den offentlige sektor og 53.000 kr. i den private sektor. I det følgende analyserer vi disse sektorspecifikke lønforskelle.

TABEL 1: KARAKTERISTIKA FOR CAND. POLIT'ER I FULDTIDSANSÆTTELSE ANDET ÅR EFTER DIMISSION

	Alle		Offentlig sektor		Privat sektor	
	Mænd	Kvinder	Mænd	Kvinder	Mænd	Kvinder
Lønindkomst (1.000 kr.)	491	432	416	398	532	479
Privatansatte	0,65	0,42	0,00	0,00	1,00	1,00
I finansiel sektor	0,29	0,18	0,06	0,05	0,42	0,37
I vidensservice	0,13	0,08	0,03	0,02	0,18	0,16
Karaktergennemsnit	8,11	7,86	7,91	7,81	8,21	7,74
Studiearbejde (årsværk)	2,44	2,15	2,25	2,13	2,54	2,18
Studietid (år)	7,07	6,98	7,33	7,09	6,94	6,83
Har børn	0,21	0,17	0,24	0,20	0,20	0,14
Forældre har LVU	0,28	0,26	0,26	0,26	0,30	0,26
Observationer	2.003	912	708	529	1.295	383

Note: Alle oplysninger er opgjort i andet år efter dimission. Lønindkomsten er opgjort i 2015-priser. *Forældre har LVU* er en indikator, som er lig 1, hvis mindst én forældre har en lang videregående uddannelse.
Kilde: Egne beregninger på baggrund af registerdata fra Danmarks Statistik.

Tabellen viser også, at mandlige og kvindelige dimittender ansat i den offentlige sektor har meget ens karakteristika, i gennemsnit. Mændene har med et gennemsnit på 7,33 år dog været på studiet lidt længere tid end kvinderne. Dette er den eneste statistisk signifikante forskel i de to gruppers gennemsnitlige karakteristika—ud over lønforskellen.

I den private sektor er en større andel af mændene ansat i den finansielle sektor og konsulentbranchen (vidensservice). Forskellene er dog ikke statistisk signifikante. Derimod er mændenes karakterer signifikant højere end kvindernes, og de har haft betragteligt mere erhvervsarbejde i studietiden. Endelig har en signifikant lavere andel af kvinderne i den private sektor børn.

Tallene understreger, at selv når vi fokuserer på en meget specifik population, der meget langt i deres livsforløb har foretaget de samme uddannelses- og sektorvalg, så er der observerbare forskelle mellem mænd og kvinder, både i gennemsnitsløn og i visse karakteristika. I næste afsnit analyserer vi betydningen af disse observerbare forskelle for de to gruppers lønfordelinger idet vi estimerer hhv. karakteristika og kønseffekter for de centrale percentiler i fordelingerne.

4. Resultater

Tabel 2 viser parameterestimater for fire regressioner. De to første er for hhv. mænd og kvinder ansat i den offentlige sektor mens de to sidste er for mænd og kvinder ansat i den private sektor. Den afhængige variabel i alle fire regressioner er den årlige lønindkomst, mens de forklarende variable er de karakteristika vi har vist i tabel 1 og

gennemgået ovenfor. De estimerede parametre er dermed "afkastet" målt i kroner for hver enhed af den angivne attribut. I regressionerne har vi udover de viste karakteristika også inkluderet årsdummier for at tage højde for vækst- og konjunktureffekter i lønningerne. Vi har udeladt dummyen for 2015 og samtidig centreret karaktergennemsnittet på 8, studiearbejdet på 2,5 årsværk, og studietiden på 5,5 år. Konstantleddet angiver dermed lønindkomsten i 2015 for en polit med det angivne køn og sektor, som har været på studiet i 5,5 år, arbejdet 2,5 årsværk i denne studietid og opnået et karaktergennemsnit på 8. Denne dimittend er ikke ansat i hverken den finansielle sektor eller i videnservice, dimittenden har ingen børn, og er ikke selv barn af en akademiker.

Parameterestimaterne i Tabel 2 viser, at disse baseline polit'er tjener lidt over gennemsnittet i den offentlige sektor, mens de tjener lidt under gennemsnittet i den private sektor. Herudover ses det, at lønningerne i den finansielle sektor i det offentlige, som ventet er markant og statistisk signifikant højere end i resten af den offentlige sektor, specielt for de mandlige ansatte, mens ansatte i videnservice i det offentlige ikke har højere lønninger. I den private sektor er der i gennemsnit højere lønninger i både den finansielle sektor og i videnservice. Her er det de relativt få kvinder ansat inden for videnservice, som oppebærer høje lønninger, i gennemsnit.

TABEL 2: PARAMETERESTIMATER FOR LINEÆRE LØNFUNKTIONER OPDELT PÅ SEKTOR OG KØN

	Offentlig sektor		Privat sektor	
	Mænd	Kvinder	Mænd	Kvinder
I finansiel sektor	82.445***	44.961***	42.546***	34.117**
	(11.221)	(12.687)	(7.972)	(10.849)
I videnservice	41.811	-19.013	36.242***	59.058**
	(25.216)	(15.063)	(9.985)	(17.971)
Karaktergennemsnit	6.202***	7.805***	17.703***	11.641***
	(1.727)	(1.590)	(2.011)	(3.081)
Studiearbejde (årsværk)	20.057***	14.808***	23.440***	29.316***
	(2.485)	(2.608)	(3.220)	(4.762)
Studietid (år)	-7.910***	0.424	-5.095	-12.923*
	(2.152)	(2.298)	(3.715)	(5.161)
Har børn	-8.521	-34.795***	-23.082*	-1.217
	(6.664)	(6.310)	(9.052)	(12.340)
Forældre har LVU	-5.957	-5.938	4.099	7.730
	(7.522)	(6.585)	(8.659)	(13.213)
Konstant	423.890***	404.678***	508.028***	471.732***
	(13.637)	(10.871)	(17.590)	(20.806)
Årsdummier	Ja	Ja	Ja	Ja
Test af årsdummier	(0.005)	(0.001)	(0.042)	(0.101)
R^2	0,276	0,279	0,165	0,269
Observationer	708	529	1.295	383

Note: Alle oplysninger er opgjort i andet år efter dimission. Lønindkomsten er opgjort i 2015-priser. Karaktergennemsnittet er centreret ved 8,0; Studiearbejde er centeret ved 2,5 årsværk; studietiden er centreret ved 5,5 år. Parameterestimater fra lineære regressioner. Robuste standardafvigelser i parentes. *, **, *** angiver statistisk signifikans på hhv. 10%, 5% og 1%.
Kilde: Egne beregninger på baggrund af registerdata fra Danmarks Statistik.

For begge sektorer og begge køn er der positive afkast af karakterer og studiearbejde mens studietiden og det at have børn "straffes" i form af lavere lønninger, i gennemsnit. Forholdet mellem karakterafkast og afkastet på studiearbejde varierer betragteligt på tværs af sektor og køn. For mænd i den offentlige sektor kan ét karakterpoint substitueres med lige under fire måneders studiearbejde, mens substitutionsforholdet er godt ni måneder i den private sektor. For kvinderne er der større ensartethed. Her er substitutionsforholdet lige under fem måneder i den private sektor og godt seks måneder i den offentlige sektor.

Det er også interessant at se, at kvinder med børn i den offentlige sektor tjener ca. 35.000 kr. mindre om året i forhold til kvinder uden børn. Endelig ses det, at lønindkomsten for dimittender, med akademikerforældre ikke er signifikant forskellig fra lønindkomsten for dimittender med ikke-akademikerforældre. Vi udelader derfor denne attribut i de følgende analyser.

Tabellerne 3 og 4 viser resultater af opdelingerne givet i ligning (7) og (8) for udvalgte percentiler. Alle resultater er for pseudo-percentiler, genereret af vægtede gennemsnit. Vægtningerne er foretaget for at sikre dimittendernes anonymitet. For percentil $q(\tau)$ er pseudo-percentilen beregnet ved

$$\breve{q}(\tau) = 0.1q(\tau - 0,02) + 0.2q(\tau - 0,01) + 0.3q(\tau) + 0.2q(\tau + 0,01) + 0.1q(\tau - 0,01)$$

Det er værd at bemærke, at der er betragtelige forskelle mellem top og bund i begge lønfordelinger i begge sektorer. I relative termer er forskellene nogenlunde ens, idet lønningerne omkring 90% percentilen er ca. 50% højere end lønningerne omkring 10% percentilen for både mænd og kvinder i den offentlige sektor, samt for kvinderne i den private sektor. For mændene i den private sektor er der en forskel på næsten 80% (godt 300.000 kr.) mellem top og bund. For den centrale IQR (0,25-0,75) varierer den relative forskel mellem 23% (kvinder i den offentlige sektor) og 32% (mænd i den private sektor).

Tabel 3 giver resultaterne for ansatte i den offentlige sektor. Lønforskellene mellem mænd og kvinder er relativt beskedne og kun marginalt signifikante (10% niveau) i den nedre ende af fordelingen. Forskellen er dog både markant—mere end 13.000 kr.—og statistisk signifikant ved medianen. Herefter er den stigende til ca. 30.000 kr. i den øvre ende af fordelingen. For alle percentiler finder vi, at effekten af forskelle i karakteristika er små og insignifikante. Kønseffekten er således på 7.000 kr. ved 10. percentil, 21.000 kr. ved 90. percentil og 16.000 kr. ved gennemsnittet. Samtidig er de to fejlkilder, balancefejlen, givet i (7) og approksimationsfejlen i (8) små og insignifikante. Vi kan således ikke afvise, at hele lønforskellen mellem mænd og kvinder i den offentlige sektor kan tilskrives kønseffekten.

I den detaljerede opdeling af kønseffekten på delelementerne i den nederste del af tabel 3 ses det at mænd "straffes" for lange studietider, mens kvinder har en "børnestraf", som det også fremgår af regressionerne i tabel 2. En børnestraf på ca. 10.000 kr. ved medianlønnen, dvs. ca. 2,5% af årslønnen, er betragtelig for ansatte i den offentlige sektor.

TABEL 3: OPDELING AF UDVALGTE PERCENTILER I LØNFORDELINGERNE FOR MÆND OG KVINDER I DEN OFFENTLIGE SEKTOR

	$q(0,10)$	$q(0,25)$	$q(0,50)$	$q(0,75)$	$q(0,90)$	Gns.
Løn, mænd	336.587	356.239	394.251	458.000	523.596	415.830
	(2.213)	(2.566)	(3.562)	(5.302)	(7.258)	(3.063)
Løn, kvinder	330.122	346.895	381.010	428.382	492.937	397.825
	(2.402)	(2.802)	(3.322)	(5.435)	(7.623)	(3.112)
Løn, kvinder, revægtet,	329.656	346.721	382.408	433.306	502.205	399.342
	(2.654)	(3.050)	(3.602)	(6.209)	(8.557)	(3.196)
Lønforskel	6.465*	9.344*	13.241**	29.618***	30.659**	18.005***
	(3.266)	(3.800)	(4.871)	(7.593)	(10.526)	(4.366)
Karakteristika	-0.128	0.128	1.106	2.503	7.093	1.464
	(1.596)	(1.989)	(2.467)	(4.163)	(5.553)	(2.561)
Kønseffekt	6.795*	9.198*	11.545*	24.104***	21.412*	16.254***
	(3.339)	(3.726)	(4.703)	(7.313)	(10.432)	(3.842)
Balancefejl	-0.338	-0.302	0.292	2.421	2.175	0.054
	(3.466)	(3.919)	(4.519)	(7.386)	(11.107)	(3.774)
Approksimationsfejl	0.136	0.320	0.298	0.590	-0.022	0.234
	(1.276)	(1.606)	(2.061)	(3.985)	(4.776)	(2.295)
Karakteristika:						
I finansiel sektor	0.083	0.134	0.577	1.373	1.580	0.643
	(0.111)	(0.174)	(0.576)	(1.382)	(1.753)	(0.649)
I videnservice	-0.059	-0.107	-0.501	-0.255	-0.550	-0.255
	(0.195)	(0.309)	(0.452)	(0.467)	(0.721)	(0.276)
Karaktergennemsnit	0.092	0.117	0.271	0.630	0.866	0.322
	(0.259)	(0.324)	(0.740)	(1.725)	(2.362)	(0.876)
Studiearbejde (årsværk)	0.772	0.977	1.178	2.270	3.370	1.780
	(0.567)	(0.702)	(0.836)	(1.595)	(2.383)	(1.221)
Studietid (år)	0.431	-0.247	0.342	0.495	2.607	0.158
	(0.666)	(0.713)	(0.761)	(1.156)	(1.858)	(0.658)
Har børn	-0.908	-0.884	-1.396	-1.925	-1.709	-1.318
	(0.676)	(0.662)	(0.975)	(1.349)	(1.282)	(0.914)
Kønseffekt:						
I finansiel sektor	0.573	1.100	1.291	3.074	0.775	1.578
	(0.395)	(0.564)	(0.736)	(1.900)	(4.528)	(0.956)
I videnservice	0.866	0.399	1.301	1.590	4.981*	1.998
	(0.730)	(0.906)	(1.010)	(1.354)	(2.190)	(1.051)
Karaktergennemsnit	0.032	0.011	-0.106	0.301	1.379	0.187
	(0.194)	(0.195)	(0.268)	(0.504)	(1.317)	(0.263)
Studiearbejde (årsværk)	-0.094	-1.032	-1.281	-1.225	-3.008	-1.381
	(0.659)	(0.806)	(0.953)	(1.452)	(2.485)	(0.909)
Studietid (år)	-7.212	-6.843	-17.155*	-26.074*	-46.621**	-14.579*
	(5.683)	(6.216)	(7.228)	(10.465)	(15.068)	(5.673)
Har børn	5.536*	5.675*	9.383***	9.971**	4.637	6.267**
	(2.189)	(2.379)	(2.742)	(3.827)	(5.602)	(2.229)
Konstant	-1.009	9.897	18.053	55.629	85.645	19.090
	(16.092)	(18.465)	(21.345)	(31.655)	(49.747)	(17.320)
Mænd	708	708	708	708	708	708
Kvinder	529	529	529	529	529	529

Note: Alle oplysninger er opgjort i andet år efter dimission. Lønindkomsten er opgjort i 2015-priser. Karaktergennemsnittet er centreret ved 8,0; Studiearbejde er centeret ved 2,5 årsværk; studietiden er centreret ved 5,5 år. Robuste standardafvigelser i parentes. *, **, *** angiver statistisk signifikans på hhv. 10%, 5% og 1%.

Kilde: Egne beregninger på baggrund af registerdata fra Danmarks Statistik.

TABEL 4: OPDELING AF UDVALGTE PERCENTILER I LØNFORDELINGERNE FOR MÆND OG KVINDER I DEN PRIVATE SEKTOR

	$q(0,10)$	$q(0,25)$	$q(0,50)$	$q(0,75)$	$q(0,90)$	Gns.
Løn, mænd	399.047	441.530	497.625	582.064	711.792	531.905
	(3.239)	(3.031)	(3.521)	(6.026)	(13.067)	(3.959)
Løn, kvinder	374.103	412.835	458.505	517.007	591.334	479.196
	(5.177)	(5.157)	(5.168)	(7.214)	(12.928)	(5.798)
Løn, kvinder, revægtet,	380.082	423.201	465.209	525.424	621.238	491.321
	(5.372)	(5.286)	(5.773)	(8.004)	(23.469)	(6.830)
Lønforskel	24.945***	28.695***	39.119***	65.057***	120.459***	52.709***
	(6.107)	(5.982)	(6.253)	(9.400)	(18.382)	(7.021)
Karakteristika	7.584*	9.712*	7.297	10.369	16.876	11.906*
	(3.677)	(3.867)	(3.867)	(5.415)	(9.415)	(4.756)
Kønseffekt	17.642**	16.894**	31.291***	55.508***	85.877***	38.955***
	(5.887)	(5.741)	(6.444)	(9.362)	(25.358)	(7.185)
Balancefejl	-1.605	0.655	-0.594	-1.953	13.028	0.218
	(6.878)	(6.942)	(7.479)	(10.172)	(25.800)	(8.394)
Approksimationsfejl	1.324	1.435	1.125	1.132	4.677	1.630
	(2.200)	(2.410)	(2.655)	(4.207)	(12.769)	(3.863)
Karakteristika:						
I finansiel sektor	0.875	0.995	0.722	1.202	1.023	0.950
	(1.147)	(1.293)	(0.962)	(1.569)	(1.475)	(1.236)
I videnservice	0.671	0.827	1.000	1.431	2.565	1.349
	(0.845)	(1.034)	(1.240)	(1.786)	(3.243)	(1.660)
Karaktergennemsnit	1.282	1.852	0.874	2.050	4.400	2.791
	(1.031)	(1.321)	(0.841)	(1.562)	(3.151)	(1.902)
Studiearbejde (årsværk)	5.098*	6.545**	7.429**	11.491**	15.585*	9.860**
	(2.041)	(2.326)	(2.500)	(3.827)	(6.341)	(3.298)
Studietid (år)	-0.885	-0.194	-0.854	-2.236	-3.365	-1.627
	(0.896)	(0.559)	(0.855)	(1.825)	(2.961)	(1.365)
Har børn	-0.082	-0.234	-0.331	-0.896	2.067	-0.118
	(1.223)	(1.174)	(1.078)	(1.151)	(2.283)	(0.850)
Kønseffekt:						
I finansiel sektor	0.188	-5.585	-0.171	-4.137	24.143	2.753
	(5.827)	(5.478)	(5.978)	(8.428)	(21.021)	(6.309)
I videnservice	-3.794	-3.190	-3.277	-2.933	-19.595	-5.436
	(2.693)	(2.793)	(3.187)	(4.923)	(13.784)	(3.772)
Karaktergennemsnit	1.061	0.672	2.172*	3.604*	1.246	1.187
	(0.798)	(0.677)	(0.970)	(1.457)	(2.798)	(0.959)
Studiearbejde (årsværk)	0.168	0.214	0.007	-0.255	-2.080	-0.317
	(0.274)	(0.285)	(0.194)	(0.402)	(2.692)	(0.421)
Studietid (år)	-5.060	-14.481	-2.083	19.658	76.022*	13.296
	(8.713)	(7.479)	(8.525)	(12.272)	(32.764)	(10.152)
Har børn	-2.109	-0.066	-1.375	-3.127	-26.712*	-3.369
	(3.771)	(3.975)	(3.860)	(4.362)	(12.971)	(3.536)
Konstant	54.286*	65.681*	51.201	-12.567	-41.856	38.637
	(25.660)	(26.086)	(30.385)	(41.952)	(110.011)	(28.887)
Mænd	1295	1295	1295	1295	1295	1295
Kvinder	383	383	383	383	383	383

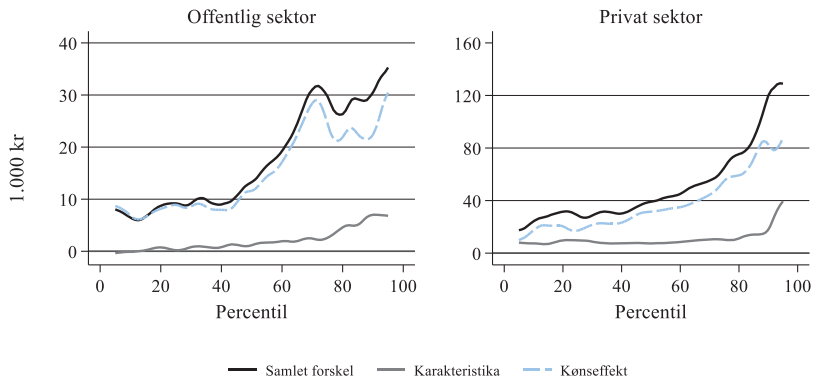
Note: Alle oplysninger er opgjort i andet år efter dimission. Lønindkomsten er opgjort i 2015-priser. Karaktergennemsnittet er centreret ved 8,0; Studiearbejde er centeret ved 2,5 årsværk; studietiden er centreret ved 5,5 år. Robuste standardafvigelser i parentes. *, **, *** angiver statistisk signifikans på hhv. 10%, 5% og 1%.

Kilde: Egne beregninger på baggrund af registerdata fra Danmarks Statistik.

Lønforskellene i den private sektor er både markante og signifikante på tværs af percentilerne, som det fremgår af tabel 4. Samlet set kan lønforskellen ikke henføres til forskelle i karakteristika, men det er værd at bemærke at den detaljerede opdeling af karakteristika viser, at mænd aflønnes for deres betydeligt længere studiearbejde, og forskellen er størst i de øvre percentiler. Kønseffekten er stor—ca. 17.000 kr. —i bunden af fordelingen, og stigende til mere end 85.000 kr. i toppen af fordelingen. Den gennemsnitlige kønsforskel er estimeret til ca. 40.000 kr. i overensstemmelse med resultatet i Eifler & Hansen (2024). Til sammenligning med den offentlige sektor, kan vi ikke genfinde en børnestraf i den private sektor. Der er derimod tegn på en børnebelønning til kvinderne. Gevinsten skal ses i lyset af at både kvinder og mænd i samplet arbejder fuld tid.

Figur 1 viser den samlede lønforskel og de rene karakteristika og kønseffekter for alle percentiler fra 5% til 95%. Figuren viser udglattede estimater (glidende gennemsnit) for at sikre anonymitet. Den stigende lønforskel over lønfordelingerne, som fremgår af tabellerne, er tydelig. Samtidig ses, at effekterne af forskelle i karakteristika overalt er små, så kønseffekten er den dominerende forskel på tværs af fordelingerne i begge sektorer.

FIGUR 1: ESTIMEREDE LØNFORSKELLE SAMT RENE KARAKTERISTIKA OG KØNSEFFEKTER OVER 5-95 PERCENTILERNE



5. Afslutning

Økonomer har længe beskæftiget sig med ligeløn mellem mænd og kvinder, typisk ved at opdele den observerede lønforskel i to; en forklaret og en uforklaret del. I teorien bør den forklarede lønforskel fange resultatet af alle karakteristika og frivillige valg, mens den resterende, og uforklarede lønforskel, fanger en systematisk forskelsbehandling på arbejdsmarkedet, som kan tilskrives individernes køn. I praksis har det været vanskeligere at opnå enighed om, hvorvidt den uforklarede lønforskel også rent faktisk skyldes

diskrimination. Et ofte hørt argument imod, at kalde den uforklarede lønforskel for diskrimination, er at der ikke er kontrolleret for *alle* forskelle i karakteristika. I vores analyse, kunne man for eksempel anføre, at forskelle i jobfunktioner eller familierelaterede forskelle kan forklare mere af lønforskellen end vi gør her. Omvendt kan metoden også kritiseres for at henføre for meget af den forklarede lønforskel til forskelle i valg, hvor det her er underforstået at valg er frivillige. Hvis kvinder for eksempel fravælger den private sektor fordi de antager, eller ved, at det vil være (for) svært at kombinere børn og karriere, kan frivilligheden i dette valg formentligt føles begrænset for de berørte kvinder og familier. Som økonomer vil vi sige, at kvinderne, og familierne, optimerer under nogle bi-betingelser, og at dette er uproblematisk. Denne økonomfaglige tilgang hjælper os dog ikke til at mindske hverken lønforskelle mellem mænd og kvinder, eller den ulige konsekvens af at få børn, som kvinder rammes af.

I den offentlige debat er interessen for lønforskelle mellem mænd og kvinder, og deres underliggende årsager, også stor. Det ses blandt andet ved, at en række danske fagforeninger siden 2010 har ført en kampagne om "Kvindernes Sidste Arbejdsdag". Kampagnen, som blandt andet støttes af økonomernes fagforening, DJØF, møder hvert år kritik for at sprede myter og misinformation fordi den ikke i tilstrækkelig grad forklarer og fortæller om, hvorfor mænd i gennemsnit tjener mere end kvinder på det danske arbejdsmarked. Diskussionen handler grundlæggende om, i hvilken grad lønforskellene kan henføres til mænd og kvinders forskellige valg i studielivet, familielivet og arbejdslivet relativt til "uforklarlige" forskelle i lønningerne.

I denne artikel har vi forsøgt at kortslutte nogle af de gængse kilder til argumentationer omkring, at uforklarede lønforskelle må skyldes karakteristika der er iboende i hhv. mænd og kvinder. Det har vi gjort ved at analysere lønforskellene for en lille gruppe personer, som har foretaget mange ens valg i studie- og arbejdslivet. Vi har således en god forståelse for, hvem de mænd og kvinder, som vi sammenligner, er. Disse personer har for eksempel udvist samme interesse og evner for matematik i deres gymnasietid, og de har oplevet den samme socialisering, som polit-uddannelsen på Københavns Universitet måtte udsætte sine studerende for. Vi betinger desuden vores analyse til, at alle økonomerne er i fuldtidsarbejde kort tid efter deres dimission. Samtidig opdeler vi økonomerne efter deres ansættelse i enten den offentlige eller den private sektor, og vi tager via regressioner højde for forskelle i deres evner og kundskaber i det omfang det afspejles i deres karaktergennemsnit på polit-uddannelsen, og deres opnåede erhvervserfaring i studietiden. Vi sammenligner dermed lønninger for en lille gruppe mænd og kvinder, som har meget ens karakteristika, bort set fra deres køn.

For at dykke lidt dybere end de ofte sete gennemsnitlige lønforskelle, ser vi på lønforskelle langs hele lønfordelingen for hhv. mænd og kvinder i hhv. den private og den offentlige sektor. Analysen viser, at lønforskellene inden for hver sektor to år efter dimission, som kan henvises til observerbare forskelle i personlige valg og evner, er meget lille—som forventet, idet vi analyserer en homogen gruppe. Størsteparten i lønforskellene er således kønseffekten.

I den offentlige sektor er den uforklarede lønforskel mellem mænd og kvinder i gennemsnit 16.000 kr. og statistisk signifikant. Ser vi på tværs af lønfordelingen, er den uforklarede lønforskel ved lønfordelingens 10. percentil 7.000 kr., mens forskellen er 21.000 kr. ved lønfordelingens 90. percentil. Både i toppen og bunden af lønfordelingen er forskellene kun marginalt signifikante. I den private sektor er den uforklarede lønforskel statistisk signifikant på tværs af hele fordelingen. Den uforklarede lønforskel er her 39.000 kr. i gennemsnit. Ved lønfordelingens 10. percentil er forskellen 17.000 kr. i den private sektor, mens den er 86.000 kr. ved lønfordelingens 90. percentil.

Forskellige i jobfunktioner eller familierelaterede forskelle kan, som nævnt, muligvis forklare en del af kønseffekten. Men det efterlader spørgsmålet hvorfor mandlige og kvindelige økonomer, som har foretaget så mange ens valg i deres formative år, begynder at divergere i deres valg kort tid efter dimissionen?

Referencer

Blau, F. D., & Kahn, L. M. (2017). The Gender Wage Gap: Extent, Trends, and Explanations. *Journal of Economic Literature*, *55*(3), 789–865.

Blinder, A. S. (1973). Wage Discrimination: Reduced Form and Structural Estimates. *The Journal of Human Resources*, *8*(4), 436–455.

Eifler, N., & Hansen, H. (2024). Lige løn for lige uddannelse? Cand.polit'ers køn og løn. *Samfundsøkonomen*, *3*, 14–23.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*(1), 259–276.

Firpo, S., Fortin, N., & Lemieux, T. (2018). Decomposing Wage Distributions Using Recentered Influence Function Regressions. *Econometrics*, *6*(2), 28.

Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica*, *77*(3), 953–973.

Hampel, F. R. (1974). The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, *69*(346), 383–393.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, *30*(1), 239–267.

Kunze, A. (2017). The gender wage gap in developed countries. In The Oxford Handbook of Women and the Economy (pp. 369–394). Oxford University Press.

Larsen, M., Verner, M. & Mikkelsen, C.H. (2020). Den uforklarede del af forskellen mellem kvinders og mænds timeløn. VIVE.

Lechner, M. (1999). Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification. *Journal of Business & Economic Statistics*, 17(1), 74–90.

Oaxaca, R. (1973). Male-Female Wage Differentials in Urban Labor Markets. *International Economic Review*, *14*(3), 693–709.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.