Symposium i anvendt statistik 2023

Økonomisk Institut
Copenhagen Business School

Symposium i
Anvendt Statistik

SYMPOSIUM I ANVENDT STATISTIK

2023

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

Støttet af SAS Institute Inc.

Økonomisk Institut, CBS 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, Copenhagen Business School, vært for symposiet, hvilket vi gerne vil takke for. Symposiet arrangeres af Symposium i Anvendt Statistik og Økonomisk Institut, Copenhagen Business School. 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 videnskabelige indlæg, er bidragsyderne ansvarlige for indholdet af indlæggene, og spørgsmål herom kan rettes direkte til forfatterne.

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

Peter Linde, Organisationskomiteen

ISBN 978-87-989370-3-6

Trykt hos PRinfoTrekroner i 110 eksemplarer

Organisationskomiteen for Symposium i Anvendt Statistik 2023

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

Peter Linde
Statistisk konsulent
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
Matematiske Fag
Aalborg Universitet
Fredrik Bajers Vej 7
9220 Aalborg Ø
esben@math.aau.dk

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

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

Helle M. Sommer
SEGES Innovation
Landbrug & Fødevarer
Axeltorv 3
1609 København V
hmso@seges.dk

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

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

Klaus Rostgaard
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

Sara Armandi
SAS Institute
Købmagergade 7-9
1050 København K
Sara.Armandi@sas.com

Birthe Lykke Thomsen
Afdeling for Børn og Unge, Rigshospitalet
Blegdamsvej 9
2100 København Ø
btho0101@regionh.dk

Hans Bay
Arbejdstilsynet
Landskronagade 33
2100 København Ø
hba@at.dk

Arne Henningsen
Fødevare- og Ressourceøkonomi
Københavns Universitet
Rolighedsvej 25
1958 Frederiksberg
arne@ifro.ku.dk

Indholdsfortegnelse

Statistisk metode

Økonomi

Statistisk analyse

The 'how' and 'when' questions for the effect of a flipped classroom inspired RCT intervention

Julie Buhl-Wiggers, Dep. of Economics, CBS
Lisbeth la Cour, Dep. of Economics, CBS
Annemette Kjærgaard, Dep. of Management, Society and Communication, CBS

1. Introduction.

As a university teacher you care about the learning of your students. Your time is also scarce so when considering introduction of new pedagogical formats, you would like to know if they are promising and actually help increasing the learning of your students. The Flipped Classroom (FC) idea is theoretically promising. Hence both quantitative and qualitative studies have been called for in the literature to assess the effectiveness of FC in higher education. In this study we contribute with an assessment of the effects of an RCT of a flipped classroom intervention in a first year, second semester macroeconomics course at CBS in the spring of 2018. Our focus on the 'before-COVID19' semester of 2018 is of relevance, as we here had time to set up our RCT taking the pedagogical ideas into account. We have analyzed both the overall 'path' from our intervention and to the final exam grade as well as investigated potential heterogeneity and mechanisms. In this paper we present a selection of our results with main focus on the 'How' (mechanisms - mediation) and 'When' (heterogeneity - moderation) questions. The average treatment effects (ATE) are briefly touched upon as well. To summarize, our research questions (RQ) are:

1. Does a FC inspired intervention on average lead to increases in student's grades (the "if" question)?
2. Do variables that relate to student personality broadly defined moderate the effects investigated in RQ1 (the "when" question)?
3. Does student class attendance work as a mediator for the effect of our FC intervention on the student's grade, when also considering potential additional controls and moderators from RQ2 (the "how" and also 'when' question)?

2. Briefly on the existing literature.

In general, flipped classroom appears to positively impact academic achievement in higher education, see e.g. Strelan et al. (2020)). The idea is that FC works because of the increased activity and engagement of students in-class (Fisher et al., 2018). However, when zooming in on the effects of FC on academic achievement in business and economics education much less agreement is found. Still some studies find positive effects (e.g. Calimeris & Sauer, 2015), some find no effects (e.g. Setren et al. (2021)) and some find both positive and negative effects depending of the prior academic achievement of the students (e.g. Asarta and Schmidt, 2017). So far not many large scale RCT's have been conducted to actually assess the magnitude of the effect of introducing FC in

cases where the traditional elements to increase active learning do not work. Furthermore, in the literature the results are based on a mixture of qualitative and quantitative studies with a majority of studies being of a qualitative nature. To our knowledge not many papers have investigated for one thing mechanisms (mediation) but also heterogeneity (moderation) quantitatively, and especially not their combination. For a single study on mechanisms based on a large scale RCT, see Authors, in review (2022).

Hence with the present study we contribute to the literature on effects of FC by providing answers to our three RQs based on a carefully designed intervention that was implemented by a - for the present context - large scale RCT.

3. The setting of the RCT and balancing tables.

In the RCT, the focus is on the tutorial classes of the course as this is where students are supposed to work actively with the problem sets. The lectures are the same for both the treated and the control group and outside the scope of the RCT. The focus on the tutorial classes is based on the observation that even though students are supposed to take actively part in the tutorial classes, usually they show up un-prepared and are quite passive during these classes. The pedagogically informed intervention has the aim of increasing activity during the tutorial classes to ultimate increase student learning and exam performance. The idea behind this intervention is inspired by the philosophy of a flipped classroom although it is not strictly speaking a classical example of FC due to the focus on the tutorial classes and not on the lectures.

The flow in learning activities during a typical week in the macroeconomics course is described below. Keep in mind that the lecturing teacher is the same for all students irrespectively of their group belongings (treated/control):

Treated class: Students are supposed to read the text of the week before the lecture (90 minutes). Students work on solving practice questions at home using MyEconLab (practice with immediate feed-back). Next they are supposed to show up in the tutorial class to work on and discuss the `main problem set' of the week. Video solutions for the main problem set are made available afterwards.

Control class: Students are supposed to read the text of the week before the lecture (90 minutes). Students work on the `main problem set' at home. In the tutorial class, the solution to the `main problem set' is presented by the teacher. Students are also offered the same extra set of practice questions as provided to the treated students. These questions are provided as a pdf-file and not through MyEconLab. Students in the control group do not have access to video solutions of the `main problem set'.

Notice that all students are provided with the same problem sets, however in different formats.

To ensure that the conditions - apart from the main idea of the intervention - was as similar as possible we further made sure that each tutorial teacher taught both a control and a treated class (to be able to remove any teacher effects). Also no time-of-the-day or day-of-the-week effects could be expected as tutorial classes were placed on the same

week day and back-to-back for each teacher with a reversing of the order in the middle of the course. Finally, a student research assistant was placed by the entrance to each tutorial class to make sure that only students assigned to this class was entering.

Table 3.1: Balance of observable controls.

		Control		Treated	p-value
Variable	N	Mean/SE	N	Mean/SE	
Age	192	21.740	223	21.758	0.721
		[0.144]		[0.092]	
Female (dummy=1 for female)	192	0.406	223	0.354	0.697
		[0.036]		[0.032]	
Capital region, Denmark	192	0.464	223	0.507	0.544
		[0.036]		[0.034]	
Non-capital region, Denmark	192	0.385	223	0.377	0.939
		[0.035]		[0.033]	
Foreign (mostly Norwegians)	192	0.151	223	0.117	0.436
		[0.026]		[0.022]	
Grade microeconomics	192	5.589	223	5.489	0.650
		[0.242]		[0.224]	
High School GPA	192	9.044	223	9.021	0.777
		[0.095]		[0.083]	
Index of self-efficacy	129	30.186	129	30.543	0.734
		[0.475]		[0.424]	
Reply dummy (dummy=1 if student took the personality test)	192	0.708	223	0.704	0.979
		[0.033]		[0.031]	
Motivation	130	3.731	131	3.664	0.531
		[0.090]		[0.087]	
Conscientiousness	136	46.463	157	45.293	0.174
		[0.556]		[0.522]	

Note: The p-value is the p-value of the corresponding t-test. In all the regressions we also include the strata variables: age, gender and high school location.

The randomization took place at the individual student level. The randomization was stratified based on the pre-determined variables of age, gender and high school location.

Students were informed about the RCT before the start of the semester and were told up front that they could ask for their data not being used in the study. (Only one student asked for this). The project has been approved by the CBS ethical committee.

All teachers have prior to semester start been participating in workshops on FC and facilitation of collaborative, active learning offered by the Teach&Learn unit at CBS.

Due to data cleaning, the final estimation sample was smaller than the starting sample and we therefore show the balancing of the two groups based on observable variables in Table 3.1. As seen in Table 3.1, there are no significant differences between the control and treatment groups of students. Hence we assume, as is standard in the literature, that this is also the case for unobservable variables.

Turning now to our hypotheses about causality:

Overall, we expect the intervention (= treatment) to affect the exam grade positively due to increased activity during the tutorial classes of the treated students.

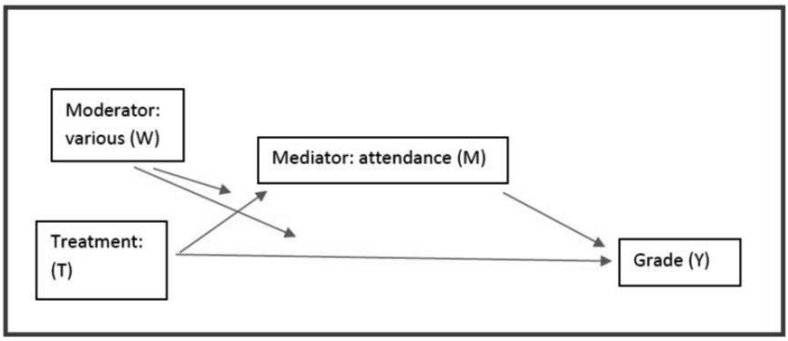
Second, we are aware that mediating effects may be present (the 'how' - mechanism). We expect that student class participation (measured by class attendance) is affected by the treatment and that class participation affects the exam outcome. Some students may like the new format some may not.

Finally, we are also aware of potential heterogeneous effects (the 'when'). We expect that several factors can work as moderators: previous student academic outcomes (proxies for ability), motivation, self-efficacy, personal traits and maybe also a behavioral variable measured as a dummy for whether or not the student filled-in the form for personality traits (the reply-dummy). In the end we only use the motivation and reply-dummy as controls and/or moderators when it comes to the extended mediation analyses following Hayes (2018). More on this later.

The models we have in mind can be illustrated by Figure 3.1, which is inspired from Hayes (2018), figure D, page 441.

In some of our estimations we apply all the arrows (translated to regressions) while in others we only focus on a subset. In some regressions we add the potential moderators just as controls before introducing the actual moderating effects.

Figure 3.1: Potential causal paths.



Note: Moderator: the 'when'; Mediator: the 'how'. We also add control variables to the model equations. Expected causal links are illustrated by the arrows. The letters in parentheses will be used for notation in the suggested regressions.

4. The data.

Our main data comes from the administrative student data of CBS. Here we have access to student background data like age (in years), gender (female = 1), high school location (capital region, outside capital region and foreign), high school GPA and also data that relates to a student's adaptation to the university environment: the grade in the microeconomics course of their first semester. In the first week of the semester during the lecture class, we ran a small survey and from that survey we have information about whether a student was looking forward to the course. We use this information as a control and also as a moderator in the analysis. Also, we asked the students to fill-in a survey to allow us to derive a measure of self-efficacy and also performed a test for measuring student personality traits (Big Five: conscientiousness, openness, extraversion, agreeableness, neuroticism). Both survey and test were given to the students prior to them being informed about whether they belonged to the treatment group or not. As a potential mediating variable, we consider class attendance (share of tutorials that a student showed up for) as we have no missing values for this variable. Without showing up in class a student would not benefit from the collaborative part of the active learning intervention.

4.1 Data pre-processing

To clean the data we dropped observations if 1) students who did not begin their studies in the fall of 2017, 2) students who dropped out during the experiment, 3) students who despite our efforts belonged to the control group but had gained access to the on-line material (only a few), 4) students who did not have a grade in microeconomics and 5) students who did not receive a grade in the macroeconomics course at the ordinary exam.

This means that the overall sample drops from 644 to 415 students. As seen above, this drop in number of observations still leaves us with balanced samples for the control and treatment groups when considering observables. For samples that are merged with variables based on surveys, the size can be considerably smaller and we have to keep that in mind through the analyses.

5. Statistical methodology.

Our analyses mainly rely on traditional regression analysis and OLS estimation. However, in order to cope with mediation and moderation, we need to also work with interaction terms (for moderators) and combinations of equations (mediation).

Equations (1) and (2) below show the most comprehensive models of the analysis. The simpler models are sub models of these. We rely on conditional process analysis as described in Hayes (2018). The concepts from Hayes (2018) consist of three types of effects from T (Treatment) to Y (Outcome=grade) when potential mediation is allowed: the *total effect*, the *direct effect* and the *indirect effect*. The *total effect* comes from an equation that does not take mediation into account:

EQ (1): $Y_i = a_0 + a_1T_i + a_2W_i + a_3W_iT_i + \text{controls} + e_i$

The subscripts 'i' refer to student 'i', T is 'treatment' and Y is the grade in macro. The W variable is for the 'base' of a moderator (differs along the way) and the interaction term W*T captures the moderation effect. Controls consist of age, gender and high school location. For balanced samples controls are actually not needed but adding them may increase the precision of the estimation. We also include a set of teacher dummies.

The *average treatment effect* (ATE) is measured by a_1 from equation (1) and without W and the interaction term in the model.

With moderation, the total effect is the effect of T on Y conditional on W:

The *total effect* is $(a_1 + a_3W)$ from EQ (1).

When mediation is taken into account, we need two more equations: one for the effect of from treatment (T) to the mediator (M) and one from M to Y conditional on T. We specify these equations as:

EQ (2): $Y_i = b_0 + b_1T_i + \gamma M_i + b_2W_i + b_3W_iT_i + \text{controls} + v_i$

Here M is the mediator, class attendance. This equation captures the effect of M on Y conditional on T.

The *direct effect* is $(b_1 + b_3W)$ from EQ (2).

EQ (3): $M_i = c_0 + c_1T_i + c_2W_i + c_3W_iT_i + \text{controls} + u_i$

The *indirect effect* is the product of γ from EQ(2) and $(c_1 + c_3W)$ from EQ (3).

The error terms, e, v, and u are each assumed to fullfill the usual assumption of OLS. We are aware of a potential danger of heteroscedasticity and we address this by applying robust standard errors when possible.

To calculate the standard errors of the coefficients that enter the expression for the indirect effect, a bootstrap method (with resampling at the class level) is applied.

With the present specification the models ensure that the *total effect* equals the sum of the *direct effect* and the *indirect effect*. For models with both moderation and mediation this holds true for the intercept and slope parts of these effects as well.

With no moderator, these effect-expressions simplify a lot (by removing all terms with W).

6. Descriptive statistics.

The first row of Table 6.1 shows descriptive statistics for the outcome variable. In the middle part between the horizontal lines, we show the mean and standard deviations for what we consider pre-determined variables. The last row of the table is for class attendance, a variable from the time period where the intervention took place.

Later, during analysis, we apply a standardized measure of the macro grade to follow the tradition in the literature. This transformation does not affect the significance of our results. For the moderation analysis we create interaction terms based on de-meaned variables for continuous moderators. Grade variables on the right-hand-side of our equations are not standardized. Notice that for the motivation variable (from initial survey) and conscientiousness (one of Big Five personality traits; from the initial test) we observer smaller sample sizes. We also have measures of the other personality traits but as it turns out that this sample causes problems later on, we just show the descriptive statistics for this single trait in Table 6.1. We assume that this trait may be the most relevant one for the present analysis.

Table 6.1: Descriptive statistics

VARIABLES	Number of observation	Mean	Standard deviation
Grade in macro	415	5.318	3.807
High School GPA	415	9.032	1.275
Grade in microeconomics	415	5.535	3.345
Index of self-efficacy	258	30.364	5.110
Female (dummy=1 for female)	415	0.378	0.486
Age	415	21.749	1.690
Capital region, Denmark	415	0.487	0.500
Non-capital region, Denmark	415	0.381	0.486
Foreign (mostly Norwegian)	415	0.133	0.339
Conscientiousness	293	45.836	6.529
Reply dummy (= 1 if student took personality test)	415	0.706	0.456
Motivation	261	3.697	1.010
Class attendance (share of tutorial classes attended)	415	0.549	0.277

Note: The grades are measured on the Danish 7 point scale. The regions are dummy variables.

7. Estimation results of the regression models

In this section we report our estimation results. We start by presenting the *average treatment effect* of the intervention. Then we move on to discuss heterogeneity and finally we turn to investigation of a potential mechanism: mediation.

7.1 Average treatment effects (the 'if' question) and sample issues

In Table 7.1 we report the *average treatment effects* estimated based on a version of equation (1) without W-terms and with and without controls. The first column shows the 'pure' and *total effect* of the intervention. The effect of the intervention is positive but insignificant at traditional significance levels. The size of the effect is rather small– around 0.10 standard deviations (a move from the 25^{th} to the 50^{th} percentile of the grade distribution is equivalent to 0.52 SDs) – although Kraft (2018) call such an effects a 'middle' sized one. In the second column we add individual background variables as

controls. The size and sign of the *average treatment effect* does not change much which is also what we would expect when working with balanced groups. The coefficient is still not significant for a double-sided test at traditional significance levels. If we apply a single-sided test the effect becomes significant at the 10% level, though. To prepare for the moderation analyses that comes next, we also look at the regression results for samples corresponding to the use of potential moderating variables. Here our focus is still on variables that we see as pre-determined or proxies for pre-determined student characteristics – but of a more personal nature. Focusing on a sample that includes the motivation variable (column 3) does not seem to lead to any bias for the estimation of the effect of the intervention. The same holds true for a sample based on availability of our self-efficacy measure (column 4). However, for the sample that allows use of the personal trait variable conscientiousness, we see a quite dramatic change in the estimated treatment effect (column 5). Hence working with the sample that allows inclusion of this (and maybe other personal trait) variable(s) seems to be a bad idea.

The results of Table 7.1 lead us to suggest the motivation and the self-efficacy variables for moderation, but not the personal trait of conscientiousness. Instead, we will try to apply the variable that somehow relates to the personality of the students but at the same time does not lead to a loss of observations, the reply-dummy: the dummy for whether a student replies to the test that is the basis of the construction of our personality trait variables. This dummy equal one if a student took the test. We would expect that this dummy is positively correlated to e.g. the conscientiousness variable but cannot do the calculations as the latter is missing for non-repliers.

7.2 Results from moderation analysis (the 'when' question)

We now focus on a range of potentially moderating variables all of which we assume pre-determined. In EQ (1) it means that we add one of these variables at a time as 'W'. All continuous moderating variables are de-meaned prior to these analyses. This is done to increase interpretability. The first group of variables relates to previous academic achievements and consists of the variables: grade in microeconomics and high school GPA. As we do have this information for all students in the main, full sample we do not have any sample change issues in these regressions. The second group of variables relates to personal traits or subjective features of each student and consists of: The motivation variable, the self-efficacy variable, and the reply-variable. For two of the three variables in the second group, we face the sample issues from Table 7.1. Only the last of these variables is available for the full sample.

Table 7.1 ATE for the macro grade and a check for sample dependence

VARIABLES	Full sample	Full sample	Motivation sample	Self.eff. sample	Per. trait Sample
Treatment	0.096	0.102	0.091	0.096	0.001
	(0.097)	(0.069)	(0.089)	(0.089)	(0.082)
Age		0.011	0.017	0.022	0.017
		(0.020)	(0.020)	(0.022)	(0.023)
Female		-0.236**	-0.185*	-0.196*	-0.199*
		(0.075)	(0.090)	(0.091)	(0.088)
Non-capital region		0.046	0.168+	0.168+	0.082
		(0.074)	(0.093)	(0.094)	(0.088)
Foreign		0.193+	0.037	0.067	0.178
		(0.111)	(0.138)	(0.139)	(0.129)
Grade microeconomics		0.177**	0.182**	0.179**	0.172**
		(0.011)	(0.013)	(0.013)	(0.013)
High School GPA		0.147**	0.131**	0.134**	0.160**
		(0.028)	(0.035)	(0.036)	(0.031)
Observations	415	415	261	258	293
Fixed effects	No	Teacher	Teacher	Teacher	Teacher
R-square adj.	0.000	0.498	0.522	0.517	0.514
F test stat	0.973	79.54	53.98	53.06	52.47
p-value of F	0.325	0.000	0.000	0.000	0.000

*Note: Robust standard errors in parentheses. ** p<0.01, * p<0.05, + p<0.1*

From Table 7.2 we notice that none of the interaction terms including previous academic achievement are significant. Such results also hold for the index of self-efficacy and for the motivation variable – although the motivation variable itself is significant at the 5% level. For the reply variable, we do find an interaction effect that is significant at the 10% level. Also the variable itself is significant. Due to inclusion of this significant variable, also the base effect of the treatment is changed and significant. The interpretation of the results is that the effect of the treatment differs depending on whether the student took the personality test or not. For a student who took the test, the treatment effect will be around zero ($a_1 + a_3$ with W=1 from EQ (1)). For a student who did not take the PT test the effect equals a_1 from EQ (1) which in this case is 0.316 and significant at the 5% level. The big question is now what the PT variable actually measures or how it can be interpreted. We will discuss this later.

Table 7.2 Heterogeneous treatment effects. Regressions with moderators. The 'when' question.

VARIABLES	Grade	Grade	Grade	Grade	Grade
Treatment	0.103	0.102	0.093	0.099	0.316*
	(0.069)	(0.069)	(0.090)	(0.087)	(0.130)
Grade microeconomics	0.169**	0.177**	0.175**	0.175**	0.175**
	(0.016)	(0.011)	(0.013)	(0.013)	(0.011)
High School GPA	0.147**	0.160**	0.132**	0.126**	0.148**
	(0.028)	(0.038)	(0.036)	(0.036)	(0.028)
Treat*micro	0.015				
	(0.020)				
Treat*HS-GPA		-0.027			
		(0.050)			
Self-efficacy			0.008		
			(0.011)		
Treat*Self-efficacy			0.009		
			(0.017)		
Motivation				0.120*	
				(0.060)	
Treat*Motivation				-0.000	
				(0.084)	
Reply-dummy					0.252*
					(0.111)
Treat*reply-D					-0.304+
					(0.157)
Observations	415	415	258	261	415
Other controls	YES	YES	YES	YES	YES
Fixed effects	Teacher	Teacher	Teacher	Teacher	Teacher
R-squared adj.	0.497	0.497	0.517	0.533	0.502
F test stat	71.16	69.65	43.30	49.36	63.96
p-value of F	0.000	0.000	0.000	0.000	0.000

*Note: Robust standard errors in parentheses. ** p<0.01, * p<0.05, + p<0.1. The other controls consists of Age and the gender and region dummies.*

7.3 Results from a couple of mediation analyses and a combination with moderation (the 'how' question).

In this section our estimations rely on versions of EQ (2) and EQ (3). These will be extended with first a 'W' in both equations and finally with both 'W' and 'T*W' to allow for combined moderation and mediation.

Table 7.3 Mediation and moderation analysis

Conditional Process Analysis	Coefficient	SE	t-value	p-value
Motivation as control in all EQs				
Total Effect	0.099	0.087	1.142	0.255
Direct Effect	0.169	0.088	1.921	0.056
Indirect Effect	-0.070	0.030	-2.349	0.019
Reply-dummy as control in all EQs				
Total Effect	0.102	0.069	1.479	0.140
Direct Effect	0.162	0.072	2.242	0.025
Indirect Effect	-0.060	0.024	-2.469	0.014
Reply-dummy as moderator				
Total intercept	0.316	0.130	2.427	0.016
Total slope	-0.304	0.157	-1.945	0.052
Direct intercept	0.373	0.129	2.896	0.020
Direct slope	-0.301	0.154	-1.953	0.052
Indirect intercept	-0.057	0.027	-2.103	0.036
Indirect slope	-0.004	0.025	-0.151	0.880

Note: based on versions of equation (1) and (2). The new controls and also the moderation has been added to both equations. All variables that relates to the indirect effect have (because they are constructed as a product of terms from two different equations) bootstrapped standard errors.

In Table 7.3, in the panel for the motivation variable, we analyze mediation by the attendance variable and with motivation as a control but without any moderation. The *total effect* here is close to the ATE we found in Table 7.1, column 3. This effect is insignificant. The *direct effect* is positive and larger than the total effect and has a p-value of 5.6%. The *indirect effect* is negative (as treated students has a lower tutorial class attendance than non-treated students) and significant at the 5% level. For the reply dummy as a control in all equations, we find quite similar results with coefficient sizes being very close to those of the motivation variable. The significance has changed slightly as now both the *direct* and *indirect effects* are significant at the 5% level and the *total effect* has a p-value of 0.14. Finally, we ran a version of the model with the reply dummy as a moderator as well – and in all equations. We notice that all coefficients except for the slope of the indirect effect are significant at the 10% level and some even at the 5% level. The actual calculation of the effects becomes a bit more complicated as

they all now depend on whether the student replied to the personality test or not. Translating the coefficients from Table 7.3 into effects, we see the following picture:

Table 7.4 Effects for a model with both moderation and mediation:

	For a student who replied	For a student who did not reply
Total effect	0.012	0.316*
Direct effect	0.072	0.373*
Indirect effect	-0.061*	-0.057*

*Note: ** p<0.01, * p<0.05, + p<0.1*

Significance for the no-reply students is as in Table 7.3 for the intercept terms. For the reply-students only the indirect effect is significant and at the 5% level. The magnitude of this effect is very close to the one for the no-reply students. This is in line with the result from Table 7.3 that the slope of the indirect effect is insignificant such that in fact there is one 'common' indirect effect for the reply and no-reply students. For students who did not reply the total effect becomes very small. For such students the direct effect is also rather small but larger than the total effect and the indirect effect is quite small and negative. For students who did not reply the total effect is large, positive and significant at the 5% level. The direct effect is even larger and significant at the 5% level. Finally, the indirect effect is significant at the 5% level and negative. The slope coefficients are significant except for the indirect effect and this implies that total and direct effects can be considered different for replying and non-replying students.

8. Discussion related to the interpretation of the reply-dummy variable.

As there seem to be moderation going on when we use the reply-dummy variable as a moderator, the following question now arises: what does the reply-variable actually capture and what are the pro's and con's of using it? Our initial hypothesis was that maybe this variable would capture one or more of a students' personal traits maybe in a kind of combination. We therefore ran regressions (Table 8.1) of the reply-variable on some of our other student-related variables. Due to perfect multi-collinearity, we cannot do such regressions for the individual personality trait variables or the sets or subsets of them. Raw correlations amongst the reply-dummy and the variables of Table 8.1 never exceed 32%. These are available from the authors upon request.

Table 8.1. Correlations by regressions for the reply-dummy with other student level variables

VARIABLES	(1) Reply-dummy	(2) Reply-dummy	(3) Reply-dummy	(4) Reply-dummy	(5) Reply-dummy
Age	-0.006	-0.008	0.009	0.006	0.008
	(0.013)	(0.012)	(0.012)	(0.012)	(0.011)
Female	0.094*	0.085+	0.067	0.078	0.068
	(0.046)	(0.045)	(0.055)	(0.054)	(0.056)
Non-Capital	0.042	0.038	0.068	0.066	0.066
	(0.049)	(0.049)	(0.057)	(0.057)	(0.057)
Foreign	0.004	0.002	-0.105	-0.128	-0.105
	(0.069)	(0.069)	(0.100)	(0.097)	(0.098)
Micro-grade	0.011	0.006	0.005	0.003	0.001
	(0.007)	(0.007)	(0.009)	(0.009)	(0.009)
HS GPA	-0.000	-0.003	0.021	0.015	0.016
	(0.020)	(0.020)	(0.025)	(0.025)	(0.025)
Attendance		0.228**			0.098
		(0.086)			(0.110)
Self-efficacy			-0.001		-0.003
			(0.005)		(0.006)
Motivation				0.049*	0.050+
				(0.024)	(0.026)
Observations	415	415	258	261	258
Controls	YES	YES	YES	YES	YES
Fixed effects	Teacher	Teacher	Teacher	Teacher	Teacher
R-square adj.	0.00517	0.0200	0.00229	0.0184	0.0116
F stat	1.329	2.154	0.989	1.711	1.431
p-value of F	0.243	0.0374	0.440	0.107	0.175

*Robust standard errors in parentheses, ** p<0.01, * p<0.05, + p<0.1*

In general we do not find much significance in Table 8.1. However notice that both attendance and motivation are positively significant at the 1% and 5% levels respectively. As pointed out earlier, we may think about the reply-variable more qualitatively as relating to single or combinations of personal traits of the students. A hypothesis relating to students' individual personality traits could be that the more conscientious students were also the ones taking the test and hence the use of the reply-dummy would work as a proxy for this trait and still allow us to use the full sample. In this case, we would expect these students to also show up in tutorial classes and to be more motivated. A hypothesis for such conscientious students could be that they would do well in the

course irrespectively of whether they were treated or not. This is also what we see from the estimation results as a treated student who replied to the personality test does not do any better than a corresponding one from the control group. The interesting results if this interpretation hold is, that a student who did not respond to the personality test – a less conscientious one – actually has both a quite large and significantly positive effect of the treatment (the *direct effect* of Table 7.4). This closes our discussion of potential explanations and interpretations for the reply-variable in this paper but the discussion may not be exhausted.

Clearly the technical advantage of using the reply-dummy is that we do not have to consider biased results due to sample size changes. The problem is that we are not – despite our attempt in the present case - able to come up with a good interpretation for this variable even though it appears significant in the combined mediation with moderation model. This leaves us to suggest a potential for further research for such reply-dummy variables, and not just as additional outcome variables (as in Damm et. al. (2021)) but also as proxy variables for otherwise unobservable variables.

9. Summary and conclusion.

We find a positive effect of the FC intervention as a reply to RQ1, although it is quite weak. We then focus on variables that could potentially lead to heterogeneous effects (RQ2). Of these we mainly concentrate on the motivation variable and the reply-dummy variable (which measures whether a student initially has taken the personality test they were offered). We find significance of motivation as an additional control variable but not as driving heterogeneity. For the reply dummy we find indications both for being a control and for moderation. Finally, when considering the self-efficacy variable, we do not find significance neither for being a control nor for moderation. Turning to RQ3, when we condition on student class attendance, we find a larger and more significant effect - the effect that we call the *direct effect* from the mediation analysis - both when the moderation variable and the reply dummy are used as additional controls. The reply-dummy is also significant as a moderator for the mediation analysis. We discuss potential interpretation of the reply-dummy variable and suggest further research for the application of such variables.

10. References

Asarta, C. J., & Schmidt, J. R. (2017). Comparing student performance in blended and traditional courses: Does prior academic achievement matter? *The Internet and Higher Education, 32*, 29–38.

Authors (2022): Insights from a randomized controlled trial of flipped, *In Review*

Calimeris, L., & Sauer, K. M. (2015). Flipping out about the flip: All hype or is there hope? *International Review of Economics Education, 20*, 13–28. https://doi.org/10.1016/j.iree.2015.08.001

Damm, A.P., Mattana, E., Nielsen, H.S. and Rouland, B. (2021), Academic achievement and wellbeing of dual language learners: Evidence from a busing program, *Journal of Urban Economics* 126, pp 1-26.

Fisher, R., Perényi, Á., & Birdthistle, N. (2018). The positive relationship between flipped and blended learning and student engagement, performance and satisfaction. *Active Learning in Higher Education*, 146978741880170.

Foldnes, N. (2016). The flipped classroom and cooperative learning: Evidence from a randomised experiment. *Active Learning in Higher Education*, *17*(1), 39–49.

Hayes, A.F. (2018), *Introduction to Mediation, Moderation and Conditional Process Analysis. A Regression-based Approach*, second edition, Guildford.

Kraft, M.A. (2018), Interpreting Effect Sizes of Education Interventions. *Brown University. Working Paper.*

Setren, E., Greenberg, K., Moore, O., & Yankovich, M. (2021). Effects of Flipped Classroom Instruction: Evidence from a Randomized Trial. *Education Finance and Policy*, *16*(3), 363–387.

Strelan, P., Osborn, A., & Palmer, E. (2020). The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels. *Educational Research Review*, *30*, 100314. https://doi.org/10.1016/j.edurev.2020.100314

Fra kaos til læring? I Covid-19's slipstrøm

Julie Buhl-Wiggers, CBS, jubu.eco@cbs.dk
Nils Karl Sørensen, SDU, nks@sam.sdu.dk
Sine Zambach, CBS, sz.digi@cbs.dk

Resumé
Gentagne nedlukninger af undervisningen på universiteterne har skabt store udfordringer. Med udgangspunkt i spørgeskemabaserede evalueringer fra SDU og CBS ser vi på erfaringerne fra studerende og underviserne.

Vi finder, at de studerendes tilfredshed med undervisningen under nedlukningen var relativ, og faldt især i foråret 2021. Generelt er præ-optagede videoer og quizzer blevet vurderet positivt af såvel undervisere som studerende. Skal undervisningen reformeres, kræver det således en nyfortolkning af rollerne for aktørerne i læringsprocessen. Endelig findes det for fællesfaget Mikroøkonomi, at karaktererne ikke nødvendigvis faldt som følge af nedlukningen på SDU og CBS.

Indledning
Siden starten af Covid-19 pandemien har undervisere og studerende lagt en stor indsats i at designe og bruge online læringsaktiviteter. Selvom vi nu er vendt tilbage til en mere normal hverdag, er onlineundervisningen i et eller andet omfang formentlig kommet for at blive. Allerede inden Covid-19 pandemiens udbrud blev digitale læringsredskaber anset som en del af fremtiden for universiteterne, især i forhold til at imødegå de seneste års stigning i antallet af studerende på videregående uddannelser og den medfølgende udfordring for at skabe dialog og interaktion. Med store holdstørrelser er mulighederne for dialog begrænset, og de studerende bliver ofte passive modtagere af viden frem for aktive bidragydere til læringsprocessen. Set fra et læringsperspektiv er dette ikke optimalt, og det er blevet en væsentlig udfordring for kvaliteten af undervisningen.

Udviklingen mod højere brug af digitale redskaber blev accelereret i hidtil uset tempo, da underviserne fra den ene dag til den anden måtte omlægge til online undervisning, og vi står nu i en situation, hvor alle undervisere i et eller andet omfang har stiftet bekendtskab med digitale undervisningsredskaber. Foråret 2020 var præget af såkaldt "nødundervisning", hvor underviserne uden de fornødne didaktiske eller tekniske kompetencer måtte omlægge til online undervisning. Man kan sige, at kaos skabte innovation i brugen af virkemidler. Sidenhen har mange undervisere eksperimenteret med forskellige digitale redskaber, såsom korte videoer, quizzer, interaktive dokumenter, streaming, podcasts etc. De sidste to år har givet et væld af erfaringer, som det danske undervisningssystem ikke har haft før. Nu har mange undervisere muligheden for at vælge, hvilke redskaber de vil tage med videre i deres formidling.

I denne artikel ser vi nærmere på, hvilke digitale redskaber der har været gode erfaringer med, og hvordan de studerende har oplevet forløbet med forskellig grad af online undervisning. Udgangspunktet for resultaterne er spørgeskemabaserede evalueringer, som studerende afgiver i forbindelse med undervisningen, underviserspørgeskemaer samt karakterer i faget Mikroøkonomi. Som følge af en identisk metodik er det muligt at sammenligne resultater fra Copenhagen Business School (CBS) og Syddansk Universitet (SDU). Undersøgelsen omfatter resultater fra HA/BA-uddannelserne på de to institutioner. Valget er foretaget ud fra en betragtning om, at de to uddannelser er relativt ens.

Digitale redskaber i universitetsundervisningen – hvor var vi før Covid-19?

Inden for de seneste 10 år har vi set en acceleration i antallet af videnskabelige artikler, der beskæftiger sig med teknologi og digitale værktøjer i universitetsundervisning. Særligt har litteraturen udviklet sig fra at fokusere på fuld online undervisning og Massive Open Online Courses (MOOCs) til at fokusere mere på en kombination af online og tilstedeværelsesundervisning også kaldet blended eller hybrid learning. Meta-studier finder generelt, at der ikke er stor forskel i læringsudbyttet mellem online og tilstedeværelsesundervisning, men at læringsudbyttet er lidt højere for blended learning. Derfor bliver blended learning ofte italesat som kombinationen af det bedste fra begge verdner (Bernard et al., 2014; Means et al., 2013).

Med stigende optagelsestal og færre ressourcer per studerende kan det uddannelsespolitisk være fristende at erstatte dele af forelæsningerne med online forelæsningsvideoer. Et sådan effektiviseringsrationale har dog også været med til at indgyde en vis skepsis hos både studerende og undervisere mod øget digitalisering og online forelæsningsvideoer (Rattleff & Holm, 2009). Inddragelse af digitale læringsredskaber behøver dog ikke være forbundet med besparelser, men giver derimod mulighed for at forbedre eksisterende undervisningspraksisser f.eks. ved brug af Student Response Systems i forelæsninger eller strukturering af undervisningsmateriale via læringsplatforme. Digitale redskaber kan også muliggøre en gentænkning af undervisningsaktiviteterne f.eks. ved at inddrage aktører fra andre sektorer eller lande via video og podcasts eller skabe plads til mere aktiverende undervisningsformer og feedback. Særlig fokus har der været på det pædagogiske format, der kaldes "flipped classroom", som bygger på tanken om at frigøre tid i undervisningslokalet til mere aktiverende undervisning ved at flytte den traditionelle formidling af pensum til videoer som bruges som forberedelse til den fysiske undervisning (Bergmann & Sams, 2012).

På de danske universiteter og erhvervsakademier bliver der også eksperimenteret med digitale redskaber i undervisningen. På Erhvervsakademi Aarhus blev der fra 2014 til 2016 iværksat et forsøg med flipped classroom med henblik på at øge de studerendes motivation og læring. Konklusionen fra dette forsøg var, at ca. 50% af de studerende

oplevede, at deres læringsudbytte var steget. Dette var dog meget afhængigt af, hvordan den enkelte underviser implementerede flipped classroom (Nørgaard, 2016). På Københavns Universitet har der været stor succes med at omlægge kurset Econometrics II til at indeholde mere aktiv læring og skabe overensstemmelse mellem læringsmål, undervisning og eksamen med brug af digitale redskaber som videoer og peerfeedback (Tabor & Müllen, 2020). På Syddansk Universitet har der ligeledes været succes med at implementere flipped classroom for ingeniørstuderende, hvor der også blev fundet en statistisk signifikant læringseffekt i den endelige eksamen (Schmidt, 2014).

På CBS blev der i 2018 påbegyndt en omstrukturering af faget Makroøkonomi på HA Almen baseret på principperne i flipped classroom (Buhl-Wiggers, la Cour, & Kjærgaard, 2022). HA Almen er et studie med over 600 studerende fordelt på 14 øvelseshold. Formålet med omstruktureringen var at øge de studerendes læring ved at skabe mere aktivitet i øvelsestimerne. Det vil sige, at den traditionelle tavlegennemgang af øvelserne blev lagt over på videoer, og undervisningstiden blev i stedet brugt på, at de studerende arbejdede aktivt med øvelserne i grupper, mens underviseren agerede som facilitator for grupperne individuelt. I bestræbelse på at skabe mere kausal evidens, blev denne omstrukturering som den første i Danmark evalueret vha. af et lodtrækningsforsøg, hvor halvdelen af de studerende blev undervist efter det nye format, interventionsgruppen, og den anden halvdel efter det traditionelle format, kontrolgruppen. Derudover, underviste hver holdunderviser både et flipped classroom og et traditionelt hold, så man på den måde kunne justere for undervisereffekter. Eksperimentet kørte i to år og i begge år blev der i gennemsnit fundet en positiv, men statistisk insignifikant effekt af det nye undervisningsformat på eksamensresultaterne (Buhl-Wiggers, la Cour, Franck, et al., 2022).

De insignifikante gennemsnitseffekter dækker dog over heterogenitet i effekter, som bibringer vigtig viden. For det første, var de studerende i interventionsgruppen mindre tilbøjelige til at komme til øvelsestimerne end de studerende i kontrolgruppen. Justeres der for denne forskel i fremmøde, så blev den positive effekt større og statistik signifikant, hvilket indikerer, at interventionen var effektfuld for de studerende, der deltog aktivt, men at dette blev opvejet i gennemsnit af de studerende, der stoppede med at komme til øvelsestimerne. Ud fra interviews med de studerende, var den største grund til fravær fra øvelsestimerne, at de ikke følte de fik noget ud af gruppearbejdet og største delene foretrak den mere traditionelle undervisning (Buhl-Wiggers, la Cour, & Kjærgaard, 2022). Dette peger på, at omstrukturering af undervisningen også kræver en ændret studieteknik fra de studerende, som ikke nødvendigvis reagerer med forventet entusiasme. For det andet, viste det sig, at der var store forskelle i undervisernes effektivitet mellem de to undervisningsmetoder, hvor nogle undervisere havde en positiv interventionseffekt, mens andre havde en negativ. Dette var overraskende, da underviserne var meget homogene ud fra observerbare karakteristika og alle havde fået klare instruktioner omkring implementeringen (Buhl-Wiggers, la Cour, Franck, et

al., 2022). Hvilket tydeliggør vigtigheden af at identificere hvilke nye kompetencer det kræver at agere i et mere aktivitetsbaseret undervisningslokale.

De seneste to år er det ikke længere kun "ildsjælene", der har eksperimenteret med digitale redskaber – derimod er alle undervisere og studerende på landets universiteter blevet kastet ud i at eksperimentere på både godt og ondt. I næste afsnit ser vi nærmere på de erfaringer, der er blevet gjort med digitale redskaber under Covid-19 pandemien, samt de studerendes tilfredshed og akademiske præstationer på HA Almenuddannelserne på både CBS og SDU.

Covid-19 pandemien – et naturligt eksperiment

Påvirkningen af Covid-19 pandemien på universitetsundervisningen kan deles op i fire faser:

1. Forår 2020: Pludselig omstilling til fuld online undervisning
2. Efterår 2020: Forventet omstilling til delvis online undervisning
3. Forår 2021: Forventet omstilling til fuld online undervisning
4. Efterår 2021: Ingen restriktioner – dog ændret eksamensform

I den første fase var det ikke bare omstilling til fuld online undervisning, men også omstilling til nødundervisning, og dermed ikke en veltilrettelagt online undervisning. Selvom nogle undervisere havde erfaringer med digitale redskaber inden Covid-19 pandemien var langt de fleste helt uforberedte, og løsningerne måtte blive ad hoc baserede og improviserede. Endvidere var de studerende også socialt og fagligt afskåret fra deres studiekammerater, hvilket påvirkede deres trivsel (Jensen et al., 2020; Nielsen, 2021). Der er allerede lavet en række rapporter om nedlunkningens indvirkning på undervisningen i foråret 2020, som viser, at særligt den manglende interaktion, faldende motivation og opretholdelse af kvalitets undervisning var udfordrende (Jensen et al., 2020; Larsen et al., 2021; Misfeldt et al., 2020; Nielsen, 2021).

I den næste fase var der en omstilling til delvis online undervisning. Udover tid til omstilling af undervisningen, var det også en væsentlig forskel, at nedlukningen kun var delvis. Det betød, at den sociale isolation, som prægede den første fase, var mindre i efteråret 2020.

I den tredje fase kunne der trækkes på erfaringerne fra den første fase. Dette burde føre til et bedre, men ikke nødvendigvis optimalt resultat, da undervisningen igen var fuld online, og hverdagen stadig var præget af social isolation og begrænset interaktion mellem undervisere og studerende.

I fjerde fase var alle restriktioner ophævet, og det var op til den enkelte underviser, at beslutte i hvilken grad digitale redskaber stadig skulle være en del af undervisningen. Covid-19 pandemien prægede dog stadig slutningen af dette semester, hvor mange

eksamener blev afholdt som hjemmeeksamener. Desuden havde de studerende højere forventninger til undervisernes brug af supplerende digitale redskaber.

Disse faser kan hjælpe os til at blive klogere på, hvordan Covid-19-pandemien med forskellig grad af online undervisning påvirkede dels de studerendes tilfredshed og akademiske præstationer og dels undervisernes brug af digitale redskaber og disses fremadrettede brugbarhed. Et overblik over data brugt i denne artikel ses i Tabel 1:

Tabel 1: Datakilder

Datakilde	N (total)	Population	Data-type	Periode	Data brugt
Kursusevaluering, SDU Kursusevaluering, CBS	17.995 11.245	Studerende, HA Almen, Besvaret	Likert, 1-5	Efterår 2018 – efterår 2021	Tabel 2
Karakterer-data, SDU Karakterer-data, CBS	3.106 3.674	Studerende, HA Almen, bestået	Karakterskala, 02-12	Efterår 2018 – efterår 2021	Tabel 3
Underviser-survey Covid-19, CBS	1.189	Undervisere, Hele CBS, besvaret	Likert, 1-5	Forår 2020 – efterår 2021	Figur 1

Online fatigue i foråret 2021

Tabel 2 viser den samlede overordnede tilfredshed på HA Almen-studiet i perioden fra efteråret 2018 og tre år frem. Generelt vurderes CBS bedre end SDU over alle perioder, mens det forholder sig omvendt med svarprocenten, der er højest på SDU. I perioder med fuld online undervisning har svarprocenten været vigende og vurderingen lavere, hvilket meget vel skyldes den manglende kontakt mellem undervisere og studerende. Især i fase 3 i foråret 2021 havde betydelig lavere tilfredshed og svarprocent også sammenlignet med fase 1 i foråret 2020. Dette er overraskende, da man måtte formode, at undervisere og studerende havde lært af erfaringerne fra det foregående år. Samme tendens er dog observeret på Københavns Universitet, hvor de studerende udtrykker stigende utilfredshed med online formatet i løbet af foråret 2020 (Jensen et al., 2020). Værd at bemærke er, at perioderne med delvis online undervisning blev vurderet på samme niveau som inden Covid-19, dog med lavere svarprocent som potentielt kan skævvride resultatet. Der er desuden en generel tendens til at studerende evaluerer mere positivt om efteråret end om foråret.

Tabel 2: Studerendes tilfredshed på CBS og SDU på HA Almen

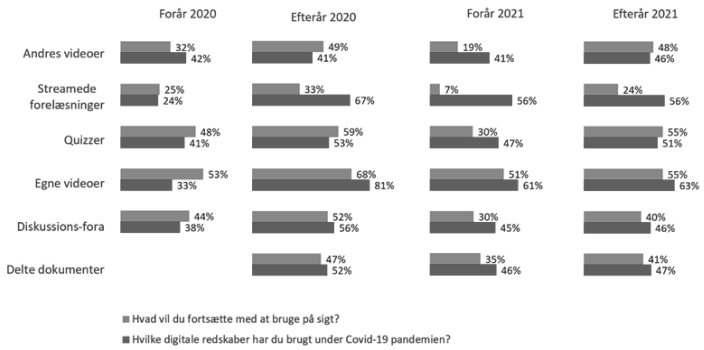
			SDU		CBS	
Periode			Resultat	Svarprocent	Resultat	Svarprocent
Fase 4	Efterår	2021	3.46	35.5	3.98	28.1
Fase 3	Forår	2021	3.29	24.8	3.56	13.6
Fase 2	Efterår	2020	3.41	34.6	3.86	26.7
Fase 1	Forår	2020	3.36	30.7	3.73	19.0
Præ Covid-19	Efterår	2019	3.69	47.7	3.94	28.3
Præ Covid-19	Forår	2019	3.42	44.6	3.76	22.7
Præ Covid-19	Efterår	2018	3.44	48.7	3.88	31.9

Noter: Egne beregninger baseret på materiale fra SDU Analytics og for CBS: Business Information and Analytics. Svarskala er 1-5 Likert skala. For CBS gælder det at: Forår 2021 er statistisk signifikant lavere end forår 2020 og forår 2019, samt at efteråret 2020 er statistisk signifikant lavere end både efterår 2021, efterår 2019 og 2018 (t-test, alfa=0,05).

Blandt underviserne var præ-optagede videoer mest populære

I fase 1, på CBS, indførte man hurtigt streaming-værktøjer, videooptagersoftware samt øget brug af Learning management-systemets funktioner, som quizzer, diskussionsfora samt interaktive dokumenter til fælles brainstorm, m.m. I fase 2, var breakout-rooms blevet state of the art, og dette muliggjorde summe-grupper, øvelsesgrupper og lignende. På SDU var udviklingen meget tilsvarende. Løbende blev poll/quiz-værktøjer som Slido og Menti introduceret, og der var mindre webinars fra sommer/tidligt efterår 2020 på CBS med introduktion til de forskellige redskaber. Da fase 3 startede i foråret 2021, var værktøjskassen derfor rimeligt opgraderet og klar til online undervisning, som igen blev påkrævet. På SDU var der en tilsvarende værktøjskasse. Endelig kan man sige, at der i fase 4, som var en "back to normal" var muligheder for at variere med de nyudviklede digitale redskaber og f.eks. gøre øget brug af videoer, quizzes og lignende.

Figur 1 viser udviklingen i brug af digitale redskaber over de fire faser, samt forventningen om brug af disse redskaber i det lange løb.



Figur 1: Undervisernes brug af digitale redskaber under Covid-19 pandemien og i fremtiden på CBS

Noter: Datakilde er Covid-19 experience project på CBS (Zambach & Kjærgaard, 2022). Svarprocenter er hhv. 60%, 35%, 28% og 25%. I foråret 2020 var spørgsmålene: "What types of online tools did you have experience with in teaching prior to the COVID-19 lockdown?" og "What will you continue with using on the long run, once the COVID-19 restrictions have been lifted?" Fra Efterår 2020 og frem blev det første spørgsmål ændret til: "What type of digital tools have you been using in online teaching last year under several levels of COVID-19 restrictions?". Der var ingen ændringer til det andet spørgsmål.

Figur 1 viser, at der er en høj brug af streamede forelæsninger og optagede videoer af en selv eller andre gennem de fire faser. Forventningen om, hvad man regner at fortsætte med, er varierende. Selvom over 50% af underviserne brugte streaming under Covid-19, tilkendegiver betydeligt færre, at de vil fortsætte med at bruge denne mulighed. Til gengæld vil over 50% fortsætte med at bruge egne videoer og quizzer. Videoerne kan bruges som indspark til undervisningen og quizzerne kan give en mere motiveret tilgang til læringen. For eksempel ses det i efteråret 2020 at hele 81% af underviserne dette semester havde gjort brug af videoer, mens 68% ville fortsætte med at bruge dem. Forventningen ligger i nogle af undersøgelserne højere end den brug der har været under Covid-19 pandemien og kan tages som et udtryk for, at underviserne i stigende grad ønsker at nuancere og variere formidlingen. Da tendensen er særligt høj for foråret 2020, kan det også være påvirket af at spørgsmålet oprindeligt spurgte ind til, hvilke redskaber der blev brugt *før* Covid-19 pandemien. Som med de studerende ses der en online fatigue hos underviserne i foråret 2021, hvor vurderingen af fremtidig brug af digitale redskaber er markant lavere end de andre semestre. Kun mellem 4 og 10 procent vil dog ikke bruge nogle digitale redskaber fremover.

Spørgeskemaerne viser også, at undervisernes indstilling til fuld online undervisning er faldende både i fase 3 og 4 dvs. foråret og efteråret 2021. Omvendt er indstillingen til

blended learning stigende (Zambach & Kjærgaard, 2022). Samme tendens er fundet i en undersøgelse fra Aarhus Universitet (Godsk, 2021). Dette tyder på, at mange undervisere har haft positive erfaringer med specielt videoer og quizzer, og at disse i et vist omfang vil blive taget med videre i deres post-Covid-19 undervisning. Blandt de studerende var særligt videoerne populære under den fulde nedlukning, mens diskussionsfora, og streaming var de mindst populære (Nielsen, 2021).

Pandemien var hård for de studerende, men de akademiske resultater var stabile

Covid-19 nedlukningen og omstillingen til online undervisning kan påvirke de studerendes akademiske præstationer både direkte via dårligere og mindre undervisning og indirekte via dårligere trivsel eller mindre motivation.

Tidligere analyser har vist, at de studerende døjede med ensomhed, manglende motivation og anså deres læringsudbytte som lavere under nedlukningen (Jensen et al., 2020; Misfeldt et al., 2020; Nielsen, 2021). Data fra UddannelsesZoom, der udføres af Uddannelses og Forskningsministeriet, viser samme tendens i både 2020 og 2021 for både SDU og CBS[i]. Vi ved dog ikke så meget om, hvordan deres faktiske karakterer blev påvirket. Tabel 3, viser karaktergennemsnittene og dumpeprocenterne for Mikroøkonomi på HA Almen for hhv. SDU og CBS. Overordnet set ser karaktererne meget stabile ud over hele tidsperioden. For SDU er der en lille stigning i karaktergennemsnittene mellem 2019 og 2020, hvor der på CBS er et lille fald, hvor 2018 dog ligger noget lavere end både 2019 og 2020. En ting, der er vigtig at bemærke er, at ikke kun undervisningen, men også prøveformen, der gik fra at være uden hjælpemidler til at være hjemmeeksamen med hjælpemidler, blev lavet om som konsekvens af Covid-19. Derfor kan vi ikke udelukke, at der også har været en ændring af sværhedsgraden af eksamen og en ændring i bedømmernes vurdering af besvarelserne, som underviserne selv bedømmer til en anelse mildere end normal (Zambach & Kjærgaard, 2022).

Tabel 3. Karakterer for Mikroøkonomi på HA Almen

		SDU		CBS	
Periode		Resultat	Dumpeprocent	Resultat	Dumpeprocent
Fase 4	2021	6.7	29	6.5	26
Fase 2	2020	6.8	23	6.4	29
Præ Covid-19	2019	6.2	26	6.8	32
Præ Covid-19	2018	6.4	27	5.9	33

Noter: Egne beregninger baseret på materiale fra SDU Analytics og for CBS: Business Information and Analytics. For CBS gælder det at karaktererne i 2018 er signifikant forskellige fra 2019, 2020 og 2021 (t-test, alfa=0,05).

På trods af *ad hoc* løsninger og forskellige institutionelle vilkår ser vi stort set de samme tendenser mellem SDU og CBS. De studerendes tilfredshed faldt betydeligt i foråret 2021, mens tilfredsheden i perioder med kun delvis nedlukning var sammenlignelig med tidligere niveauer. På hverken SDU eller CBS ser vi nogen betydelig ændring i de faktiske karakterer i Mikroøkonomi.

Afrunding og udblik: Muligheder i Covid-19's digitale slipstrøm

Covid-19 pandemien har på den ene side givet et gevaldigt digitalt kompetenceløft, men på den anden side også revet tæppet væk under både undervisere og studerende og dermed sat et afskrækkende aftryk i manges bevidsthed. Trods dette ser vi også tendenser til en mere positiv holdning til at kombinere digitale redskaber med tilstedeværelsesundervisning.

Hvis man spørger underviserne og de studerende, så er især de præ-optagede videoer og quizzer, som med fordel kunne fortsætte efter Covid-19 pandemien. Dog viser underviser spørgeskemaerne også at flere føler et stærkt pres fra de studerende, om at fortsætte med at bruge digitale værktøjer, streame og optage forelæsningerne (Zambach & Kjærgaard, 2022). Hvilket giver anledning til nye forventningsafstemninger mellem universiteterne, underviserne og de studerende samt diskussioner omkring datasikkerhed, rettigheder etc.

De præ-optagede videoer giver mulighed for at bruge undervisningslokalet mere pædagogisk og samtidig give de studerende den fleksibilitet, som de i stigende grad efterspørger. Netop samspillet med den øvrige undervisning er vigtig for at præ-optagede videoer bliver en meningsfuld læringsressource (Mathiasen, 2019). Forskning indenfor økonomiundervisning viser dog at tilgange som flipped classroom ikke altid giver den forventede læringseffekt (Buhl-Wiggers, la Cour, & Kjærgaard, 2022; Craft & Linask, 2020; Ficano, 2019; Wozny et al., 2018) og at udformningen af den fysiske undervisning har stor betydning for størrelsen af læringsudbyttet (Buhl-Wiggers, la Cour, Franck, et al., 2022; Foldnes, 2016). Da de fleste undervisere primært har erfaring med forelæsninger, kræver det en mental omstilling samt kompetence-opbygning til at agere i undervisningslokalet, hvis forelæsning ikke længere er den primære undervisningsmetode. Foruden underviserne kræver det også en omstilling af de studerendes studiestrategier, samt nyfortolkning af roller og ansvar, hvis transformering fra forelæsninger til aktivitetsbaseret undervisning skal lykkes.

Fremtidige muligheder for brug af præ-optagede videoer kunne også være en øget deling af undervisningsressourcer undervisere i mellem. En sådan deling kan både lette arbejdet for den enkelte underviser med omstilling til mere aktiv undervisning, og samtidig skabe mulighed for peer feedback. Dog er det også vigtig med refleksion over, hvad der sker med undervisning, når det bevæger sig fra at være en serviceydelse, som

konsumeres samtidig med at det produceres, til at være et fysisk produkt, som kan deles og vurderes udenfor undervisningslokalet.

Covid-19 pandemien har sat øget fokus på opkvalificering af undervisernes digitale kompetencer, men for at erfaringerne fra Covid-19 pandemiens kaos kan transformeres til varige forbedringer af de studerendes læring kræver det kontinuert pædagogisk udvikling, robuste effektevalueringer samt investeringer.

Referencer

Bergmann, J., & Sams, A. (2012). *Flip Your Classroom: Reach Every Student in Every Class Every Day*. International Society for Technology in Education.

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education*, *26*(1), 87–122. https://doi.org/10.1007/s12528-013-9077-3

Buhl-Wiggers, J., la Cour, L., Franck, M. S., & Kjærgaard, A. (2022). Investigating effects of teachers and peers in flipped classroom: An RCT study of classroom level heterogeneity. *Working Paper*.

Buhl-Wiggers, J., la Cour, L., & Kjærgaard, A. (2022). Impact of flipped classroom on academic achievement: Unforseen challenges to students' willingness to participate. *Working Paper*.

Craft, E., & Linask, M. (2020). Learning effects of the flipped classroom in a principles of microeconomics course. *The Journal of Economic Education*, *51*(1), 1–18.

Ficano, C. K. C. (2019). Identifying differential benefits from a flipped-group pedagogy in introductory microeconomics. *International Review of Economics Education*, *30*, 100143.
https://doi.org/10.1016/j.iree.2018.07.002

Foldnes, N. (2016). The flipped classroom and cooperative learning: Evidence from a randomised experiment. *Active Learning in Higher Education*, *17*(1), 39–49.
https://doi.org/10.1177/1469787415616726

Godsk, M. (2021). Coronapandemiens indflydelse på universitetsadjunkters holdning til teknologi i undervisningen. *Tidsskriftet Læring og Medier (LOM)*, *14*(24), Article 24.
https://doi.org/10.7146/lom.v14i24.125580

Jensen, L. X., Karstad, O. M., Mosbech, A., Vermund, M. C., & Konradsen, F. (2020). *Experiences and challenges of students during the 2020 campus lockdown. Results from student surveys at the University of Copenhagen* (pp. 1–73). University of Copenhagen.

Larsen, S., Georgsen, M. (Ed.), Qvortrup, A. (Ed.), Andersen, I. S. K., Asmussen, I. S., Bak, C. K., Buus, L., Dalsgaard, C., Geisnæs, D., Graf, S. T., Gundersen, P., Gynther, K., Horn, L. H., Jensen, S. T., Jørgensen, A., Jørnø, R., Kjærgaard, T., Konnerup, U., Larsen, I. K., Lorentzen, R. F., Lyngsø, A., … Troelsen, R. (2021). *Erfaringer og oplevelser med online undervisning på 9 videregående uddannelsesinstitutioner i foråret 2020* (M. Georgsen & A. Qvortrup, Eds.).

Mathiasen, H. (2019). Video, en læringsressource i universitetsundervisningen. *Tidsskriftet Læring og Medier (LOM)*, *12*(21), Article 21. https://doi.org/10.7146/lom.v12i21.112627

Means, B., Toyama, Y., Murphy, R., & Baki, M. (2013). The Effectiveness of Online and Blended Learning: A Meta-Analysis of the Empirical Literature. *Teachers College Record*, *115*.

Misfeldt, M., Jensen, L. X., Hvillum, N. P., Harboe, T., Lindvig, K., Ejrnæs, M., Pedersen, A., Horak, R., Eskelund, H., & Larsen, L. N. (2020). *Evaluering af online-nødundervisning forår 2020* (1st ed.).

Nielsen, M. M. (2021). *Questionnaire survey of students' experiences during the COVID-19 lockdown. Answers to open-ended questions* (pp. 1–105). Copenhagen Business School.

Nørgaard, C. (Ed.). (2016). *Flipped Classroom Muligheder og barrierer*. Erhvervsakademi Aarhus.

Rattleff, P., & Holm, L. G. (2009). Barrierer for ibrugtagning af videooptaget universitetsundervisning. *Tidsskriftet Læring og Medier (LOM)*, *2*(2), Article 2. https://doi.org/10.7146/lom.v2i2.3911

Schmidt, B. (2014). Improving motivation and learning outcome in a flipped classroom environment. *2014 International Conference on Interactive Collaborative Learning (ICL)*, 689–690. https://doi.org/10.1109/ICL.2014.7017854

Tabor, M. N., & Müllen, R. L. von. (2020). Et statistikfags succesfulde omstrukturering – fokus på alignment og god feedbackpraksis. *Dansk Universitetspædagogisk Tidsskrift*, *15*(28), 51–70.

Wozny, N., Balser, C., & Ives, D. (2018). Evaluating the flipped classroom: A randomized controlled trial. *The Journal of Economic Education*, *49*(2), 115–129.
https://doi.org/10.1080/00220485.2018.1438860

Zambach, S., & Kjærgaard, A. (2022). *Experience from COVID-19 among teachers in three semesters at Copenhagen Business School*. Copenhagen Business School. https://covid19exp.cbs.dk/.

[i] Disse resultater er ikke vist i artiklen, men kan rekvireres af forfatterne.

Absence and Completion among students in Vocational Education

Fane Groes, Department of Economics, Copenhagen Business School, Denmark

Edith Madsen, Department of Economics, Copenhagen Business School, Denmark

Tróndur M. Sandoy, Department of Economics, The University of the Faroe Islands, Faroe Island

We analyze the effect of school absence on program completion among a group of students in Vocational Education in Denmark. According to human capital theory, being present in class and participating in class activities is an important determinant of human capital formation and therefore the causal effect of absence on educational performance is of interest. To identify this effect we use data on daily student attendance from the administrative systems of VET schools in combination with register data on ompletion and student background characteristics. There is a very strong correlation between absence and ompletion. In order to identify the causal effect of absence on completion, we introduce a new panel data instrument for absence that uses variation over time in absence for the individual student. We have observations of absence over time but we only have cross-sectional information on completion. Under certain assumptions, the proposed instrument is independent of unobserved individual-specific fixed effects, such as ability, that might affect both absence and completion. However, the conditions for validity of the panel data instrument do not hold when many students have zero absence and that is the case in the setup considered in this presentation. In a different approach, we use certain weather conditions (rain and wind speed) as instruments for absence. We face the challenge that there is only little variation in weather conditions.

Keywords: Economics of Education, Vocational Education and Training, Absence, Instrumental Variables.

Textual Love

Text Analysis on Facebook Messenger Data

Sara Armandi, SAS Institute

1 Introduction

Unstructured text is the largest human-generated data source, and it grows exponentially by the minute [SAS Software, 2018]. The potential knowledge hidden in these data must not be overlooked. Unfortunately, the knowledge might seem quite inaccessible. We all know how difficult, or at least, how time consuming, it can be to retrieve insights from a single document. Now consider having to go through all medical reports in an entire hospital, all insurance case notes within a given subject or messages sent back and forth between interesting parties. This is almost impossible.

As the amount of data has increased by mind-numbing speed, luckily, so has the memory handling capabilities and hence, the computational powers of computers. Additionally, analytical procedures and techniques have been developed, which are brilliant when wanting to extract knowledge from textual data. On social media like Twitter and Facebook, text analytics is extensively used to extract patterns and detect specific behavior. This information is among other things used to help create interventions for cyber bulling, sexual predatory behavior, and even terrorism planning.

In this article, text analytics is used within a more pleasant topic. Instead of detecting cyber bullying, this article attempts to investigate cyber love. Through text analytics conducted using SAS® Visual Text Analytics, the content of 17,874 messages sent between my husband and me, using Facebook Messenger, are examined. Different love concepts, terms, topics, and categories are discussed to get insight into how much and in which ways our love can be extracted by textual analysis.

2 Theory

The amount of data generated is inconceivable. Just considering the digital footprint that we as individuals create every single day, the scope is incomprehensible. The footprint includes digital activities, actions, contributions, and communication which are manifested on the internet or on a digital device [TechHQ, 2018]. Considering the staggering amount of unstructured data that is generated every day, automated software which is rapid and consistent is critical to fully analyze text efficiently.

2.1 Text Analytics

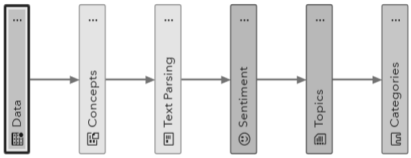
Text analytics helps analysts extract meanings, patterns, and structure in unstructured textual data [Chakraborty, 2013]. In this process, software is used to read and understand human-generated text. The software can classify, sort, and extract information from text to identify terms and concepts, sentiment, topics, and categories. The principal object in text analytics is a document.

A key algorithmic component of text analytics is natural language processing (NLP). The primary role of NLP is to efficiently parse a document collection. In semantic parsing relationships between words in a sentence is detected (e.g. it is determined whether a pronoun is the object or subject of a verb). Through this process an important list of terms is created, which both topic derivation, sentiment analysis, and text categorization depend on [SAS Institute Inc., 2021, p. 1-9].

2.2 SAS® Visual Text Analytics

SAS® Visual Text Analytics (VTA) on SAS® Viya® automatically converts unstructured data into meaningful insights. It combines the power of natural language processing, machine learning, and linguistic rules in an end-to-end analytics framework with the goal of creating insights from textual data. VTA is accessed through the SAS® Model Studio. The analytical process carried out by VTA generate four main result categories: terms and concepts obtained from a Concepts node and the Text Parsing node, sentiment from the Sentiment node, topics from the Topics node, and categories from the Categories node. The default pipeline created by VTA is shown in Figure 1. It contains all the nodes described above.

Figure 1: Default SAS® Text Analytics Pipeline



Source: Screenshot from SAS® Visual Text Analytics.

Currently, 33 languages are supported by VTA [SAS Institute Inc., 2022]. Luckily, Danish is one of these languages. Of course, English is the most evolved, but Danish do support most features. However, when running a sentiment analysis on VTA, using the Sentiment node, the following notes are produced:

```
NOTE: 2022-12-30T10:22:41.321 - Executing the task "sentiment".
NOTE: 2022-12-30T10:22:41.530 - A default sentiment model does
      not exist for the language "DANISH".
```

One could find or create a sentiment model in Danish, however, for now, this is skipped. Consequently, the sentiment analysis as well as the corresponding Sentiment node is to be discussed in a future paper.

Both the Concepts and Text Parsing nodes uses NLP. Specifically, a concept is contained within a document and can be extracted from a document. A document can have multiple concepts, and the same concept can be represented multiple times [SAS Institute Inc., 2021, p. 1-89]. Concepts are the entities, facts, events, or key pieces of information in the text data. In VTA, the concepts are parsed in the Concepts node. Nine predefined concepts are supported by the node. These include measures, persons, dates, and organizations. Custom concepts allow the user to extract business-specific information. To add custom concepts, the LITI (language interpretation for textual information) syntax, a powerful, flexible, and scalable language, is used.

The Text Parsing node parses for general language elements. The NLP tools supported by the Text Parsing node include parsing, tokenization, part-of-speech tagging, synonym detection, and stemming.

The Topics node is one of the feature extractions nodes. With this node, NLP and unsupervised machine learning helps reveal trends in data by automatically extracting terms and topics that appear in correlation to each other throughout a set of documents. This allows for quick discovery of trends in data. As an analyst, you don't even have to know explicitly what to look for ahead of time [SAS Institute Inc., 2022].

The last node to mention is a text modeling node. Categories are created by promoting topics from the Topic node to categories, by specifying a category variable when assigning roles to the project data, or by creating a new category. The node uses machine learning and NLP to derive category rules or Boolean scripts for each category variable. Rules that are automatically generated for category variables and for topics that are promoted to categories are editable. [SAS Institute Inc., 2021, p. 3-40] Finally, the Categories node allows a document to fall into more than one category.

3 Data

To get data for this analysis, I dive into my own digital footprint. On Facebook, it is possible to request and download a copy of one's profile information. When making a request, it is possible to select which types of information that should be included. Further, one can choose the date range, the media quality and the file format. The two formats are HTML

and JSON format. As the latter is stated to allow other services to import the data more easily, the JSON format is selected [Meta, 2022].

3.1 Extracting Data from Facebook Messenger

As an initial attempt, I retrieved only part of my digital footprint on Facebook. The request was made on December 1st, 2022, at 12:06 AM and included information on "Friends and followers, Posts, Comments and reactions, Messages and more". For this analysis, only the Messages exchanged with other people on Messenger are of interest. The date range was set to "All time", implying that data ranges from the account creation date (January 5th, 2008) until the date and time of the request.

The data was created as six large zip files (1.805.128 KB, 2.562.010 KB, 1.319.230 KB, 2.561.260 KB, 2.559.432 KB and 2.557.845 KB) which were available for download. When diving into the overwhelming amount of information, only one of the zip files contained the sent messages of interest. The messages inbox folder takes up 827 MB, containing 1,894 sub folders, each representing a conversation thread with one or more participants. In this analysis conversations marked as archived og filtered, as well as message requests were excluded.

Opening a couple of the individual conversation sub folders, it is seen, that these contain a JSON message file. It is these message files that are utilized in this analysis. Besides a message file, the folder might also include audio-, files-, gifs-, photos- and videos-conversation content sub folders. Hence, the conversation content sub folders show the non-text part of the messages sent between the participants in a conversation. These could be of interest but are not investigated any further in this paper.

3.2 Facebook Encoding Issues

To import the data into a SAS environment, the SAS 9.4 (Unicode Support) application is used, as this supports the UTF-8 encoding, which is also the default encoding of the JSON files. Not using the Unicode version of SAS will result in errors like:

```
ERROR: Some character data was lost during transcoding in the dataset
    XXX. Either the data contains characters that are not
    representable in the new encoding or truncation occurred during
    transcoding.
```

The messages contain a lot of characters which are difficult to handle and not supported by the default WLATIN1 encoding used in "regular" SAS 9.4 applications.

Even when using the Unicode supported application, which allows for multilingual computing, there are still a lot of characters which cannot be displayed in a proper way.

Apparently, Facebook download data is incorrectly encoded as the original data is UTF-8 encoded but was decoded as Latin-1 instead [Stack Overflow, 2018]. A mojibake ï¿½!

The largest problems occur in connection with emojis. Emojis are very frequently used in the messages sent using Facebook Messenger. To fully understand the meaning and the sentiment of messages, emojis are important. The unprocessed data doesn't show which emoji is sent, just that "a" emoji is sent. Unfortunately, this does not help a lot, as emojis can differ quite a lot. In further analysis, one could fix the bad encoding. However, in this analysis the bad encoding is disregarded and hence, the emojis are not considered.

In the context of emojis, it should also be considered that the development and the use of emojis has changed a lot over time [Emoji Timeline, 2022]. Additionally, the way people use emojis might differ quite a lot [BuzzFeed, 2022].

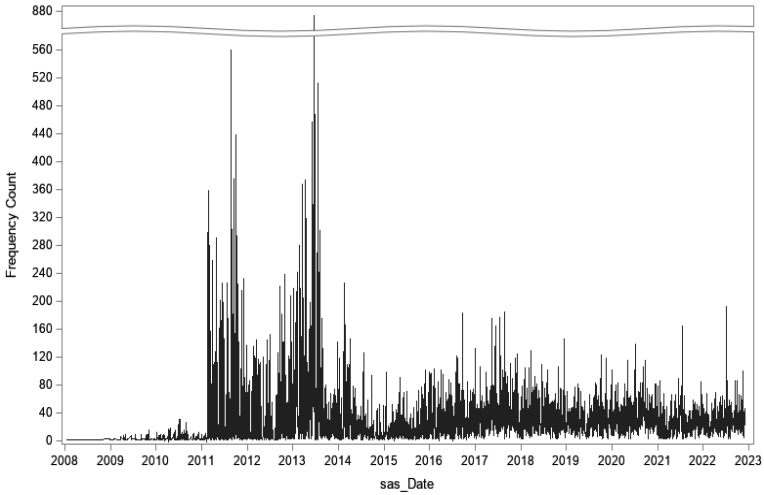
3.3 The Large Dataset

Disregarding the encoding issues, the JSON data is imported to SAS. The combined dataset contains 143,271 observations. Hence, 143,271 messages have been sent in the period from January 14th, 2008, to November 30th, 2022. The first message was sent at 10:19 PM by my high school friend Caroline Hansson, and the last was sent at 10:22 PM by my brother-in-law, Esben Scriver Andersen.

The 143,271 messages are divided on 1,894 different conversations. Some conversations are one-to-one (me and one other person messaging to each other), others are group conversations. In total, there are 1,370 unique sender names represented in data. One of these senders is of course me, accounting for 65,252 of the messages, corresponding to 45.61 percent of all messages. My husband has sent 7,949 messages, which is 5.56 percent of the total number of messages, and hence, the person out of all 1,370, who has sent most messages to me.

Figure 2 shows the number of messages sent by date. It is clear, that Facebook Messenger was only sporadically used until 2011. In the years between 2011 and 2015, the messages exploded. Especially on June 22nd, 2013, (cf. Table 2 in Section 3.4) the number of messages reached a maximum. After 2015 the number of messages has found a relatively stable level.

Figure 2: Frequency of Daily Number of Messages Sent (and Received)



Source: Output from PROC SGPLOT with SERIES statement, SAS® 9.4.
Note: There is a break in the y-axis from 580 860 in order to compress the view of the line plot.

When looking at the numbers in

Table 1, the observed picture in

Figure 2 is reinforced. The maximal number of messages sent during the 2011-2015 period catches one's eyes – 874 messages sent on a single day. It is worth noting, that regular text messages (SMS) were way more common in the early years (of the ones considered in this analysis). Now, Facebook Messenger covers almost the entire need for textual messaging. However, now and then, this changes.

Table 1: Descriptive Statistics of Daily Number of Messages Sent

Years	Number of Days	Mean	Std Dev	Min	Max
2008-2011	328	4.09	4.38	1	31
2011-2015	1,262	44.91	71.00	1	874
2015-2023	2,826	30.17	22.49	1	192
All	4,416	32,44	43.27	1	874

Source: Output from PROC MEANS, SAS® 9.4.

Note: The *Number of Days* column shows the days where at least one message has been sent.

In the dataset, there are quite a few blank observations. These observations correspond to messages which might contain pictures, files or other of the extra conversation content discussed above in Section 3.1. These blank messages are neither deleted nor handled, but just kept as they are, as they represent some kind of interaction.

Further, all messages are grouped into different types. The most frequent type in the data is *Generic*, which accounts for almost 97 percent of the messages. These are "regular" text messages or blank observation. In total 2,320 of the messages are of the type *Call*, with the first call from Facebook Messenger being conducted on July 21st, 2013. The 2,320 observations should have been removed, but in this analysis they have not.

The dataset created is quite large, mainly due to the *Content* variable. This variable contains all the content of the messages sent. The length of the variable is in SAS set to 9,736, which is necessary, as this is the number of digits used in the longest message sent. Only in this way truncation is avoided and no text is lost. This results in all values of the *Content* taking up the space of 9,736 digits, even though, in reality, they might be blank or just contain a single or two digits. In total, the large dataset ends up taking 1,389,024 KB.

3.4 Data Sent to and From My Love

Due to the intention of this paper, only part of the data is used[1]. A scenario where the occurrence of cyber love is investigated across all conversations seems a bit risky to the analyst. Hence, the data considered is limited to the messages sent between me and my husband, Søren. It is worth noting, that Søren hasn't been my husband throughout the entire period. Further, the use of other (textual) communication forms have also been used, e.g., SMS, but in this paper, we focus on messages sent using Facebook Messenger.

Table 2: Descriptive Statistics of Message Dates

Years	Min	Max	Mean	Median	Mode

[1] Additionally, SAS® Viya for Learners, which is used to do the analysis, only allows dataset less than 100 MB to be uploaded on the platform.

All Messages	14JAN2008	30NOV2022	17JUL2016	22AUG2016	22JUN2013
Søren Messages	23DEC2013	30NOV2022	20APR2019	15NOV2019	19FEB2014

Source: Output from PROC MEANS, SAS® 9.4.
Note: The *Søren Messages* row are the dates when messages were sent to and from my husband, Søren

Considering only the private conversation between my husband and me (group conversations have been deleted), we have sent 17,874 messages back and forth. As seen in Table 2, the first message was sent on December 23rd, 20213, at 8:38 PM. The last message is form November 30th, 2022, at 3:02 PM. When comparing the mean and the median in Table 2 it seems as if the distribution of the messages is left skewed, as the median exceeds the mean. However, when plotting the daily number of messages sent the distribution is closer to uniform in the years between 2016 and 2023. In 2014 the frequency is enormous and in 2015 it is very limited (cf. Figure 3 in Section 4.1).

To access the dataset in SAS® Viya® for Learners, the dataset is divided into two chunks, so the individual datasets does not exceed the 100 MB limit. However, after uploading the dataset to the Viya platform, it is easy to recreate the original dataset with all messages sent between my husband and me. In SAS® Viya® for Learners, the SAS® Studio application is available and is used with the following code:

```
data message.final_soeren;
     set message.final_soeren_1 message.final_soeren_2;
run;

proc contents data=message.final_soeren;
run;
```

4 Results

The initial results are clear. I have sent a total of 10,585 messages to my husband in our private conversation. He has only sent 7,289 messages. There can be no doubt, I love my husband more than he loves me. In this analysis we utilize the power of text analytics, and hence, more sophisticated results are presented below.

4.1 The Concept of Love

To get insights into the text, the concept of love is created. When defining the concept, words and phrases which are associated with love are extracted. Basically, this corresponds to reading through ALL messages and highlighting each time one of the words or phrases in the list appear. Often definitions or explanations of text analytics are influenced by the environment of the person who does the analysis. In this case, I, the analyst, have quite

good insights in what is contained in the text of interest as I have written most and read all messages. Based on this knowledge combined with extensive thinking and rapid and superficial discussions with my husband, the concept is defined using words or phrases that we use to express love through messages to one another. The concept is defined using the following rules:

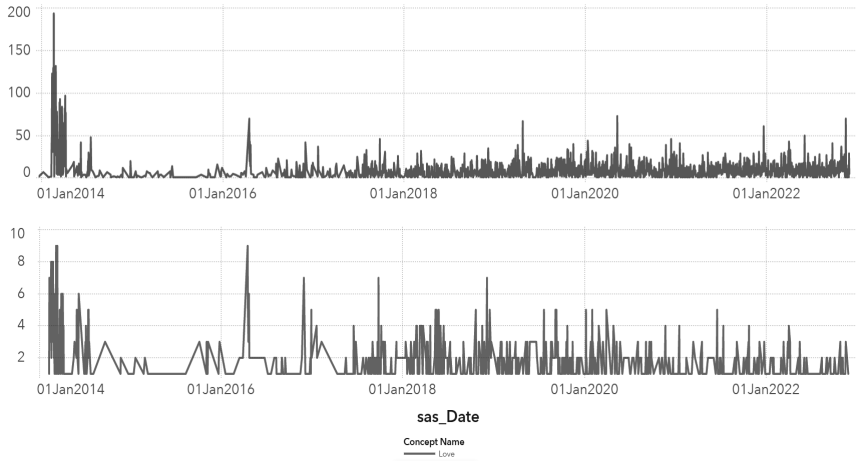
```
CONCEPT:elske@
CLASSIFIER:elskede
CLASSIFIER:love you
CLASSIFIER:amo
CLASSIFIER:quiero
CONCEPT:du er dejlig@
CLASSIFIER:dejligste
CONCEPT:du er sød@
CLASSIFIER:sødeste
CLASSIFIER:kys
CLASSIFIER:<3
```

Of the 17,874 messages, 1,450 contains at least one of the keywords from the love concept. In total, 1,790 concepts are found, resulting in an average number of matches per document of 1.23.

When looking at the individual keywords in the concepts, “elsker” is most frequently used as it appears in no less than 485 different messages. This is closely followed by the keyword “kys”, which is found in 483 messages. The phrases “du er sød” and “du er dejlig” are in total written in 80 different messages. In Figure A1 in Appendix a word cloud of the frequency of the keywords is shown. Note that the words in this plot are case sensitive. The vertical line, |, between the keywords indicate that the words appear in the same message.

To get a better understanding of the usages of the love concept, the frequency of messages is shown in Figure 3. The top row shows how all 17,874 messages are divided on the different dates. The bottom row is the frequency of messages that contains the love concept by date. The plot shows, that the amount of love is expressed in conjunction with the overall number of messages sent. In periods where more messages are sent, more love is sent as well. It is clear from Figure 3 that most love was sent in the beginning of the initial messages sent using Facebook Messenger. This period was, believe it or not, also the beginning of the relationship between me and Søren.

Figure 3: Frequency of Messages Sent by Date and Messages with Love Concept



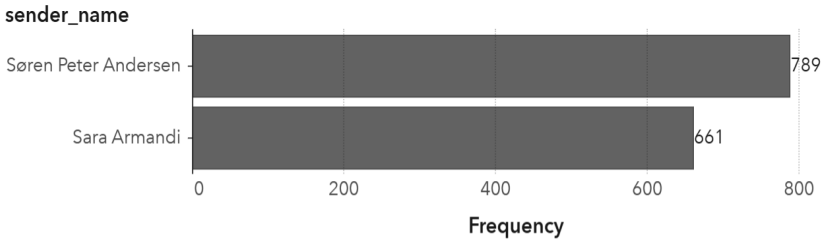
Source: Result from Transactional Output Data from Concepts node in SAS® Visual Text Analytics, processed in SAS® Visual Analytics.

Note: The scale of the y-axis differs a lot on the two plots.

As the frequency of messages has found a steady level, so has love. In the past two or three years, textual love is rarely expressed more than twice a day, and in some periods, there is no love at all.

A pressing question can not be avoided any longer: "Who loves the most". According to the results in Figure 4, the answer is clear. When dividing the 1,450 messages that contains the concept of love, I have sent only 661 of these messages, whereas my loveable husband, Søren, has sent 789. I am sure, that one reason for this surprising result is that love expressed by emojis are not considered. Had they been part of the analysis the result might have looked different (or maybe not).

Figure 4: Frequency of Love Concept by Sender

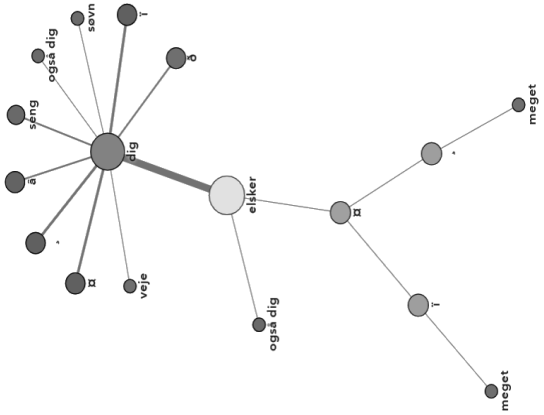


Source: Result from Transactional Output Data from Concepts node in SAS® Visual Text Analytics, processed in SAS® Visual Analytics.

4.2 The Love Term

Diving deeper into love, it is possible to investigate the different terms used in the messages. Automatically, VTA drops 10,584 from the analysis, as they are on the default stop list (words like "i, og, den, at, så, til, osv.) and are used in less than four messages. On the other hand, 2,275 terms are kept, as they seem relevant for the analysis, or if they are part of the concept rule. The most frequent term throughout all messages is "jeg" (containing both "jeg" and "mig") which is used 5,463 times in 3,878 different messages. In comparison, taking a look at one of the most love expressing terms "elsker", this is found in 483 different messages and is used no less than 488 times.

Figure 5: Term Map for Term "elsker"



Source: Result from Text Parsing node in SAS® Visual Text Analytics.

The term "elsker" is depicted in the term map above in

Figure 5. The term map is useful to identify associations between different terms. The thickness of the connector between the terms "elsker" and "dig" is an indication of the relative information gain which estimates the strength of the association. Hence, as the association is strongest between these two terms. What is interesting is that "elsker" is often associated with undefined characters. These are most likely some kind of emoji that expresses love (However, which emojis are unclear, cf. Section 3.2). Further, "elsker dig"

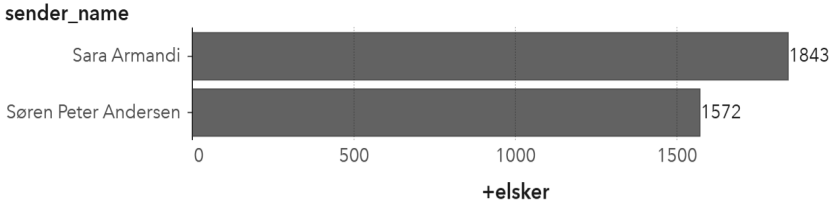
is associated with “seng” and “søvn”, terms used when it’s time to go to bed. Hence, love is expressed right before bedtime.

4.3 Love as a Topic

When considering natural groupings of important terms that are specific to particular messages, SAS® is able to automatically derive topics. A total of 14 different topics are suggested. The automate topic which is most love-like is the sixth most frequent one. This topic is related to a total of 1,384 different messages. The default name of the topic is “+dig, +elsker, kys, +jeg, +du”, which are also the five most relevant terms in the topic. The first term has a relevancy score of 0.728 and the last of the five has a relevancy of 0.104. Of the 2,275 terms which (as noted in Section 4.2 above), 34 are matched with the topic. The relevancy of the score can also be negative, meaning that these term should not appear in a message for the message to fit to the topic.

To underline, that “elsker” is one of the most important terms, this is promoted to a topic in this analysis. The new “+elsker” topic is named only by the term used to create the topic. However, the number of terms relevant for the topic is 269, with the term “elsker” having the highest relevancy score of 1.0 (as this is an exact match and the term itself). In total, the custom created “+elsker” topic is assigned to 3,415 messages. This implies that a few more messages than implied in the results above might contain love. Love that wasn’t captured by the defined love concept. These messages can be examined in more detail, as they are listed in VTA and are provided with a relevancy score. The score implies how well the message belongs to the topic.

Figure 6: Frequency of “+elsker” Tobic by Sender



Source: Result from Topics Output Data from Topics node in SAS® Visual Text Analytics, processed in SAS® Visual Analytics.

Using the new and improved love identifier, the picture as described in Section 4.1 changes. From Figure 6 it is clear, that I am suddenly the one sending more love. The picture is the same, when plotting the “+dig, +elsker, kys, +jeg, +du” topic. In this case, I am responsible

for 788 of the love messages, and my husband only accounts for 602 messages which have been assigned this love topic.

Both topics are promoted to categories. A category identifies a group of messages that share a common characteristic. The definition is based on the terms included along with its relevancy. When considering the default topic "+dig, +elsker, kys, +jeg, +du", the following Boolean rule was used to create the category:

```
(OR,(AND,(OR,"elske","elsker")),(AND,(OR,"dig","digó"),"kys"),(AND,(NOT
,"sendte"),(NOT,"sendt"),(NOT,"han"),(NOT,"ham"),(OR,"digó","dig")))
```

Now even more matched messages, in total 1,516, are found in data. Correspondingly, the category created by promoting the "+elsker" topic is also represented in way more messages, namely 4,529. The increased numbers are due to the definitions of the category rules, which are loosened compared to the topic.

When looking at who sends most love when considering the new categories, ones again, I am expressing most love in the messages I send. In the case of the "+dig, +elsker, kys, +jeg, +du" I'm sending 918 messages as opposed to 598 sent by Søren. Also when considering the "+elsker" category, I send most love with 2,621 compared to 1,908 lovely messages. One reason to why I have suddenly exceeded Søren by quite a lot is that the overall number of messages sent by me are slightly overrepresented in data. Hence this influences the text analytics models runned by VTA.

4.4 The Language of Love

It is no surprise that my husband and I have different ways of expressing ourselves in text. As the variable *sender_name*, which indicates whether a message is sent by my husband or me, is assigned the category role in VTA, the Categories node is able to provide interesting insights. For a category variable Boolean rules for each sender is derived. Investigating the script, it is possible to see how the difference in messaging is perceived by the software.

When comparing the category Boolean rules for the category definition for my husband and me, it is conspicuous how the code created to categorize messages sent by me contain only very few love related terms, cf. Note A2 in Appendix. In comparison, the script derived by the software for the category definition for my husband is given by:

```
(OR,(AND,(OR,"tigeren","tigerð","tigre","tiger")),(AND,(NOT,"uendeligt"
),(NOT,"uendelig"),(NOT,"helst"),(NOT,"helt"),(NOT,"hel"),(NOT,"¤"),(NO
T,"alle"),"kys"),(AND,"haha"),(AND,"tumling"),(AND,(OR,"mener","men")),
(AND,(OR,"igå","igår")),(AND,(OR,"kue","ku")),(AND,(NOT,"¤"),(NOT,"ð"),
(OR,"cca","ca")),(AND,(OR,"godt elskede","god elskede")),(AND,(OR,"lød"
,"lyder","lyde")),(AND,(NOT,"å"),(NOT,"ð"),(NOT,"jeg"),(NOT,"mig"),(OR,
"dejligste","dejlige","dejligt","dejligst","dejligere","dejlig"),(OR,"g
od","gode","godt")),(AND,(OR,"stress","stresse")),(AND,"jað"),(AND,"mob
```

```
il"),(AND,(NOT,"â"),(NOT,"siddet"),(NOT,"sidder"),(NOT,"siddende"),(NOT
,"sad"),(NOT,"sidde"),(NOT,"os"),(NOT,"vi"),(NOT,"ð"),(OR,"veje","vejer
","vejede","vej")),(AND,(OR,"bandit","banditter")),(AND,"pipperen"),(AN
D,(OR,"fyre","fyrene","fyr")),(AND,"je"),(AND,"ca"),(AND,(OR,"pusling",
"puslingen","puslinge","pusli")),(AND,"mindre"),(AND,(NOT,"â"),(NOT,"ð"
),(NOT,"jeg"),(NOT,"mig"),(OR,"god","gode","godt")),(AND,"fuck"),(AND,"
²"),(AND,"også dig"),(AND,"st"),(AND,"nuð"),(AND,(OR,"chicka","chick","
chickas")),(AND,(OR,"den","denð","detð")),(AND,"digð"),(AND,(OR,"smukt"
,"smuk","smukke")),(AND,"stakkels dig"),(AND,"ish"))
```

Even though this script is shorter, it contains more terms that could be related to love expressing.

The above script might show a pattern which is useful to categorize the messages when new data is retrieved. From Table 3, the precision is given by the two categories of the *sender_name* variable. Categorizing messages as if they were sent by me is a lot more precise than categorizing messages sent by my husband. Besides, that I might be more consistent in my writing, the larger precision might also be due to the fact, that messages sent by me is overrepresented in data compared to messages sent by my husband, cf. Section 3.4.

Table 3: Precision and Misclassification of Messages

Sender_Name	Precision	Number Messages	Matched Messages	Minimum Misclassified	False Negative	False Positive	True Positive
Sara Armandi	0.7572	10,585	7,635	2,950	4,804	1,854	5,781
Søren Peter Andersen	0.5763	7,289	3,484	3,805	5,281	1,476	2,008

Source: Results from Categories node in SAS® Visual Text Analytics.

Note: The column *Precision* indicates how well the category rule performs. *Number of Messages* is the number of messages actually sent. In the column *Matched Messages*, the number of messages which have been classified by the respective sender name, using the belonging Boolean rules are given. *Misclassified* is the minimum number of misclassified messages.

5 Conclusion

Love is in the air, traveling like radio-wave communications signals from one Facebook Messenger account to another. This is clear by the analysis, which shows that a total of 1,450 messages contains one or more of the concepts defined in the custom specified love concept. The love concept consists of words and phrases used by my husband and me to express love. Love was most frequently expressed (at least through Facebook Messenger text messages) in the early stages of our relationship. The last couple of years, love has found a steady level and is expressed on average in one or two messages a day.

Surprisingly, the initial analysis considering the custom created love concept shows that my husband, Søren is the best at expressing his love in text messages. Also, when looking

at the category Boolean rule for the *sender name* variable (which indicates whether a message is sent by me or my husband), my husband seems to be categorized as the one who is using most love related terms. When diving deeper into the facets of text analytics, the primary sender of love changes. According to both topics and categories created by the VTA software, I account for most messages sent with love related content.

To make sure, that the love will keep on flying back and forth in messages between my husband and me, the conclusion must be, that we both love each other – even though, overall, I have sent almost 45 percent more messages to my husband than he has sent to me. So maybe I love Søren 45 percent more than he loves me?!

References

BuzzFeed – Molly Capobianco (2022). *How Different Generations Use Emojis*. Retrieved 31 December 2022, from https://www.buzzfeed.com/mollycapobianco/different-generations-emoji-use

Chakraborty, Goutam, Murali Pagolu, and Satish Garla (2013). *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. Cary, NC: SAS Institute Inc.

Meta – Facebook Help Center (2022). *Download a copy of your information on Facebook*. Retrieved 31 December 2022, from https://www.facebook.com/help/212802592074644

The SAS Dummy – Chris Hemedinger (2015). *How to convert a Unix datetime to a SAS datetime*. Retrieved 31 December 2022, from https://blogs.sas.com/content/sasdummy/2015/04/16/how-to-convert-a-unix-datetime-to-a-sas-datetime/

SAS Help Center (2018). *Writing Concept Rules: Basic LITI Syntax*. Retrieved 31 December 2022, from https://documentation.sas.com/doc/en/ctxtcdc/8.2/ctxtug/writingconceptrules.htm

SAS Institute Inc. – Terry Woodfield and George Fernandez (2021). *SAS® Visual Text Analytics in SAS® Viya® Course Notes*. Cary, NC: SAS Institute Inc.

SAS Institute Inc. (2022). *SAS Visual Text Analytics Solutions*. Retrieved 31 December 2022, from https://www.sas.com/en_us/software/visual-text-analytics.html

Stack Overflow – Martijn Pieters (2018). *Facebook JSON badly encoded*. Retrieved 31 December 2022, from https://stackoverflow.com/questions/50008296/facebook-json-badly-encoded

SAS Software YouTube Channel (2018). *SAS Visual Text Analytics Demo*. Retrieved 31 December 2022, from https://www.youtube.com/watch?v=H30RrmGpctk

TechHQ (2018). *What do you know about your data footprint?* Retrieved 31 December 2022, from https://techhq.com/2018/11/what-do-you-know-about-your-data-footprint/

Appendix

Figure A1: Word Cloud of Love Concept Keywords

Source: Result from Transactional Output Data from Concepts node in SAS® Visual Text Analytics, processed in SAS® Visual Analytics.

Note: The words in the word cloud are case sensitive. The vertical line, |, between the keywords indicate that the words appear in the same message.

Note A2: Category code created to categorize messages sent by me

```
(OR,(AND,"¤"),(AND,(OR,"minutter","minuts","minut","minutters")),(AND,(OR,"dayoâ","dayo
")),(AND,"okay"),(AND,"ô",(OR,"jeg","mig")),(AND,":p"),(AND,(OR,"sisselô","sissel","sis
sels")),(AND,(OR,"cykel","cykle","cyklerne","cykelô","cyklen","cykler","cyklerô"),"ô"),
(AND,(NOT,"godt"),(NOT,"god"),(NOT,"gode"),(OR,"haft","har","havde","have"),"ô"),(AND,"
vedhæftet fil"),(AND,(NOT,"cca"),(NOT,"ca"),(NOT,"kys"),(OR,"forsøger","forsøgt","forsø
get","forsøgte","forsøge"),(OR,"jeg","mig")),(AND,"glip"),(AND,"hmm"),(AND,"jeg"),(AND,
"omkring"),(AND,(OR,"skate","skatô","skat","skatte")),(AND,(OR,"fhtz","fht")),(AND,"for
håbentlig"),(AND,":o"),(AND,(NOT,"god"),(NOT,"gode"),(NOT,"godt"),"ô"),(AND,"called"),(
AND,(NOT,"tigeren"),(NOT,"tiger"),(NOT,"tigerô"),(NOT,"tigre"),(NOT,"haha"),(NOT,"kl"),
(NOT,"kys"),(NOT,"tur"),(OR,"mig","jeg"),"meget"),(AND,(NOT,"tigeren"),(NOT,"tiger"),(N
OT,"tigerô"),(NOT,"tigre"),(NOT,"kl"),(NOT,"kys"),(NOT,"tur"),(OR,"mig","jeg"),(OR,"sav
net","savner","savne","savnede")),(AND,(NOT,"tiger"),(NOT,"tigre"),(NOT,"tigerô"),(NOT,
"tigeren"),(NOT,"haha"),(NOT,"kl"),(NOT,"kys"),(NOT,"tur"),(NOT,"men"),(NOT,"mener"),(O
R,"maxi","max","maxô"),(OR,"mig","jeg")),(AND,(OR,"flottest","flot","flottere","flottes
te","flotte")),(AND,"â"),(AND,"faktisk"),(AND,(OR,"formentligt","formentlige","formentl
ig")),(AND,"d."),(AND,":d"),(AND,"yes"),(AND,"igen"),(AND,(OR,"telefonen","telefon","te
lefoner")),(AND,(OR,"kørte","køre","køres","kører","kørt","kørende")),(AND,"maxâ"),(AND
,"sådan"),(AND,(OR,"cykel","cykler","cyklerne","cykle","cyklerô","cykelô","cyklen")),(A
ND,"aller"),(AND,(OR,"here","her","herô","herâ")),(AND,(OR,"videochatten","videochat"))
,(AND,(OR,"var","varer","varerne")),(AND,(NOT,"tigre"),(NOT,"tigerô"),(NOT,"tigeren"),(
NOT,"tiger"),(NOT,"haha"),(NOT,"kl"),(NOT,"cca"),(NOT,"ca"),(NOT,"kys"),(NOT,"tur"),(OR
,"mig","jeg")))
```

Faktoranalyser på mange ESS runder

Hans Bay & Anders Milhøj

hba@at.dk

anders.milhoj@econ.ku.dk

Faktoranalyse

I en faktoranalysemodel betragter man n stokastiske variable $X_1, .., X_n$ med middelværdi

$$x_i = \sum_{j=1}^{k} b_{ij} f_j + e_i$$

0, som man forsøger at forklare ved færre ($k < n$) latente faktorer $f_1, .., f_k$; eller på matrixform blot **X** = **BF** + **E**. Restleddene $e_1, .., e_n$ antages uafhængige, så de fælles faktorer i modellen forklarer al samvariationen, korrelationerne, mellem de observerede variable $X_1, .., X_n$. Modellens parametre er alle koefficienterne b_{ij} samt varianserne på restleddene. Det er nødvendigt at pålægge parametrene visse restriktioner for at modellen bliver veldefineret, fx antages at *f*-erne er standardiseret med varians 1, og at *f*-erne er uafhængige. Imidlertid er antagelsen om uafhængighed ikke nødvendig i den forstand, at det er muligt at gange og dividere med samme ortogonale matrix **A** mellem matricerne **B** og **F** i matrixligningen, uden at modellens forklaring af *X*-erne ændres. Ved at vælge et fornuftigt **A** kan man dermed opnå en bedre fortolkning af modellen ved at tillade korrelerede faktorer. Det sidste kaldes en rotation, hvor vi i dette indlæg benytter promax rotationen.

Modellens parametre kan estimeres ud fra *N* observationer af den *n*-dimensionale vektor **x** = $(x_1, .., x_n)$. Ud fra de estimerede parametre kan man estimere faktorerne, kaldet at score faktorerne, som betingede middelværdier givet de observerede **x**-er.

Faktoranalyse på tidsrækker

Hvis **x**-erne er observeret til tid $t = 1, ..,T$, kan der opstilles faktoranalysemodeller for hvert tidspunkt *t*, så modellen for X_{ti} , $i = 1 .., n_t$ og $t = 1, ..,T$ bliver

$$x_{ti} = \sum_{j=1}^{k} b_{tij} f_{tj} + e_{ti}$$

for hvert t.

I dette generelle udtryk kan de enkelte dele opfattes som tidsrækker på forskellige måder, idet dog restleddene e_{ti}, $t= 1, .., T$ for hvert i opfattes som afhængige med middelværdi 0, dvs. som en hvid støj i tidsrækkeforstand.

- Man kan estimere på det samlede datamateriale ved at opfatte datasættet som mange observationer af de n variable, dvs. negligere at de er observeret på forskellige tidspunkter. Derved kan N_t endda være 1.

- Man kan estimere faktormodellen for hvert tidspunkt, t, for sig og se, hvordan de scorede faktorer f_{tj} i modellen med ens faktorloadings $b_{tij} = b_{ij}$ udvikler sig over tid. Det kræver selvfølgelig, at antal observationer for hvert tidspunkt, N_t er stort nok, dvs. i hvert fald langt større end 1.

- De observerbare X_{ti} kan opfattes om en n dimensional tidsrække, der fx. kan modelleres ved en vektorautoregressiv model. Det kræver, at antal observationer N_t er 1. Men det negligerer jo faktoranalyserne helt.

- Faktorerne F_{tj} kan opfattes som en k dimensional tidsrække, der fx. kan modelleres ved en vektorautoregressiv model og samtidigt opfatte koefficienterne b_{tij} som konstante over tid, dvs. samme loadings hvert t. Dette er den dynamiske faktormodel, hvori N_t kan være 1.

- Koefficienterne b_{tij} kan opfattes som en nk dimensional tidsrække, der fx. kan modelleres ved en vektorautoregressiv model, men nk er ofte for stor, til at det er brugbart i praksis.

Trustvariablene for England - en faktor

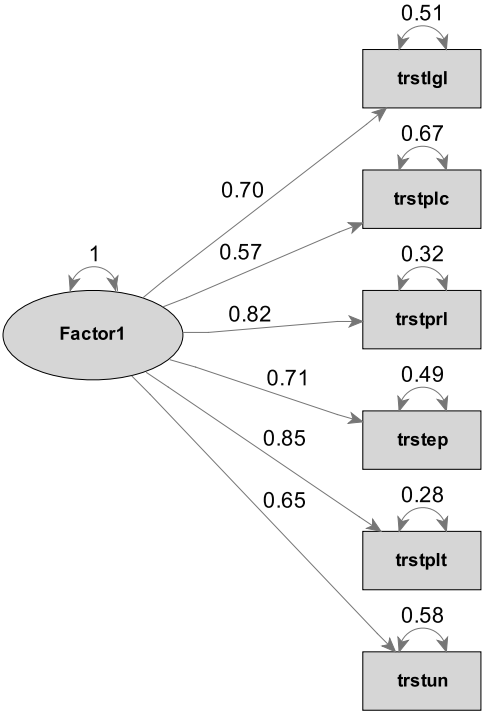
I dette eksempel anvendes 6 'trust'-variable på data indsamlet i England. Disse trust variable består af 6 spørgsmål, som alle er stillet på en skala.

No trust at all									Complete trust
1	2	3	4	5	6	7	8	9	10

Trstep:	Trust in the European Parliament
Trstlgl:	Trust in the legal system
Trstplc:	Trust in the police
Trstplt:	Trust in politicians
Trstprl:	Trust in country's parliament
Trstun:	Trust in the United Nations

Disse variable er inkluderet i runde 1 – 9 i ESS, dvs. at det er en meget kort tidsække med $T = 8$. Derimod er antal observationer pr. runde højt med N_t mellem 1586 og 2052.

En faktoranalyse med en enkelt faktor på alle trust variablene estimeret for alle runder samlet giver modellen, som er vist i Figur 1.

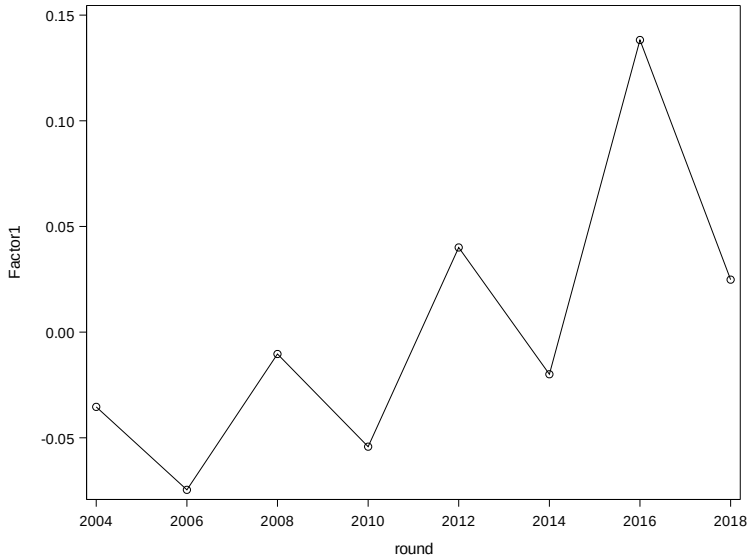


Figur 1. Faktoranalysen med en faktor

Faktormodeller med en faktor estimeret for hver runde for sig

Når faktoren scores og dens gennemsnit beregnes for hver runde for sig, bliver gennemsnittene som vist i Figur 2.

Figur 2. Den scorede faktors gennemsnit for hver runde

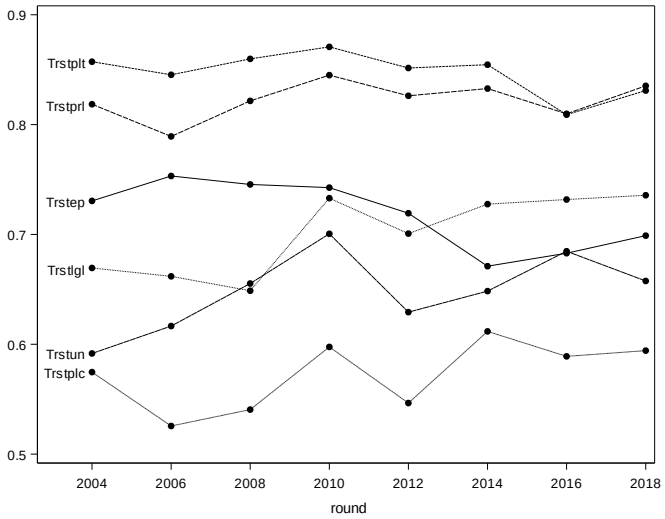


Den scorede faktor er estimeret som standardiseret med gennemsnit nul og varians en, når der estimeret for alle observationer i alle runder under et. Da gennemsnittene er beregnet for mange personer - mellem 1586 i runde 2 og 2052 i runde 4, antager gennemsnittene værdier tæt ved nul.

De er meget markant, at faktorens værdier stiger med årene, mens den takkede form tyder på en negativ autokorrelation, så tilliden går skiftevis op og ned fra runde til runde.

Ser man på faktorloadings, når faktoranalysen estimeres for hver runde for sig, bliver billedet naturligvis mere kompliceret; se udviklingen i Figur 3. Disse faktorloadings er angivet på diagrammet i Figur 1 for den samlede analyse, men de har altså ændret sig noget fra runde til runde.

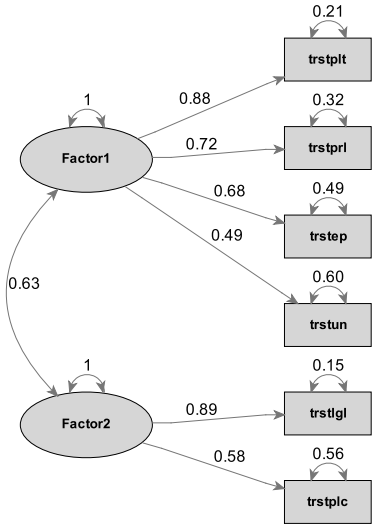
Figur 3 viser, at faktorloadings har været nogenlunde stabile over årene, så faktoranalysen virker stabil i tid. Det mest markante er, at loadings for trestep og trstplt er dalet mens loadings for trstlgl og trstprl er steget. Men det er jo ikke, fordi de pågældende variable er steget - det er snarere et mål for, hvor meget variablene hænger sammen med de øvrige variable.



Figur 3. Gennemsnit af faktorloadings for hver runde

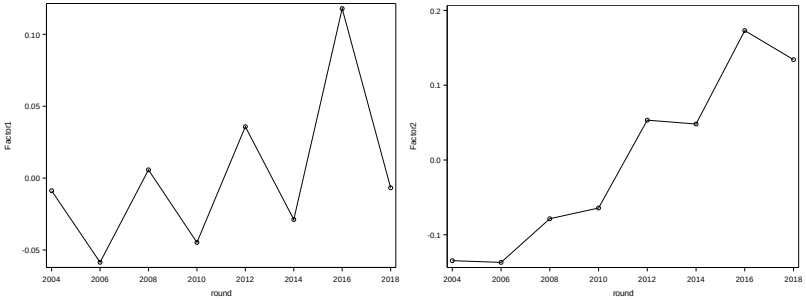
Trustvariablene for England - to faktorer

Resultatet af en faktor analyse med to faktorer på alle trust variable estimeret for alle runder er vist på Figur 4, mens Figur 5 viser gennemsnittene de to scorede faktorer for hver runde for sig af.



På Figur 5 ses, at faktor 1 udvikler sig som faktoren estimeret i modellen med kun en faktor - den er også som faktor stort set identisk med den ene faktor i en simplere mode. Faktor 2 stiger markant - det er faktoren for tiltro til politi og domstole.

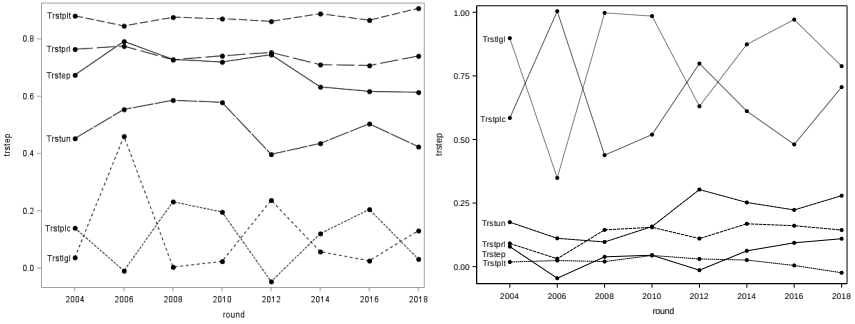
Figur 5. Faktor 1 og 2 scoret på det samlede datamateriale



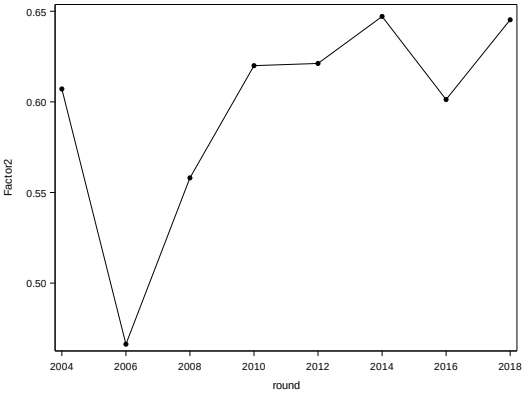
Faktormodeller med to faktorer estimeret for hver runde for sig.

Ser man på faktorloadings, når faktoranalysen estimeres for hver runde for sig, bliver billedet som vist på Figur 6. Visse år er Heywood tilfælde, så loadings antager værdien en. Figur 7 viser korrelationen mellem de to faktorer; det ses, at den dykker i 2006 med ellers har steget de senere runder.

Figur 6. Faktorloadings på faktor 1 og 2, når de scores på hver runde for sig



Figur 7. Korrelationen mellem de to faktorer



Dynamisk Faktoranalyse med en faktor

En dynamisk faktoranalyse kan estimeres med Proc SSM i SAS, men datasættet er for stort, til at PC'en kan klare det. Derfor er estimationen foretaget på en tilfældig indsamlet stikprøve på 500 respondenter fra hver ESS runde – i alt 4000 respondenter.

Modellen er lidt anderledes end i en faktoranalyse, hvilket enklest beskrives ved at stille de seks ligninger op

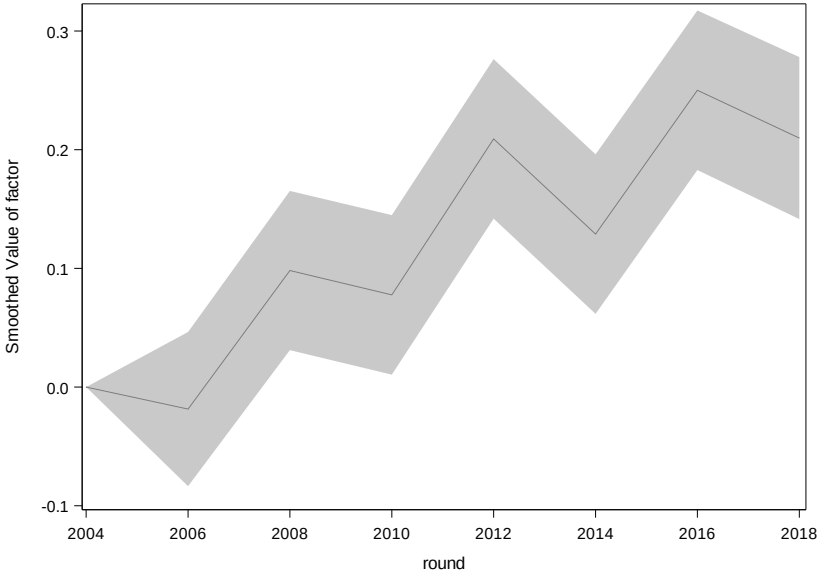
$$\texttt{trstprl} = \texttt{int1} + f + \texttt{w1};$$
$$\texttt{trstlgl} = \texttt{int2} + \beta_2 f + \texttt{w2};$$
$$\texttt{trstplc} = \texttt{int3} + \beta_3 f + \texttt{w3};$$
$$\texttt{trstplt} = \texttt{int4} + \beta_4 f + \texttt{w4};$$
$$\texttt{trstep} = \texttt{int5} + \beta_5 f + \texttt{w5};$$
$$\texttt{trstun} = \texttt{int6} + \beta_6 f + \texttt{w6};$$

I disse ligninger er `int1-int6` parametre, der lægges til den latente faktor *f*. Denne latente faktor er normeret til at have en udvikling svarende til den første variabel `trstpl`, fordi den ikke ganges med en parameter, et β_1, mens de øvrige trust-variable ganges med et β. Restleddene `w1-w6` er normale restled i regressionsmodeller.

Figur 8 viser den estimerede latente faktor estimeret som en tidsrække, der er modelleret som en hvid støj. Figuren er sådan set magen til Figur 2; men det nye er, at der er sikkerhedsgrænser, så det kan ses, at ændringerne i faktorens værdier fra runde til runde faktisk er væsentlige. Desuden er faktoren plottet ud fra et udgangspunkt på nul i 2004, så værdierne de efterfølgende år skal ses relativt til dette udgangspunkt. På nær det lille

fald fra 2004 til 2006 er der en tydelig opadgående tendens dog med takker. De seks observerede variable er forskudt i forhold til faktorens værdier med forskydninger som vist i Tabel 1.

Figur 8. En faktor estimeret som en irregulær proces



Tabel 1. Parameter estimater

Response	*Estimate*	*Std error*	*t Value*	*Pr > |t|*
trstprl	4.11	0.0445	92.30	<.0001
trstlgl	4.97	0.0801	62.09	<.0001
trstplc	5.98	0.0659	90.70	<.0001
trstplt	3.38	0.0380	89.06	<.0001
trstep	3.40	0.0376	90.35	<.0001
trstun	4.88	0.0450	108.47	<.0001

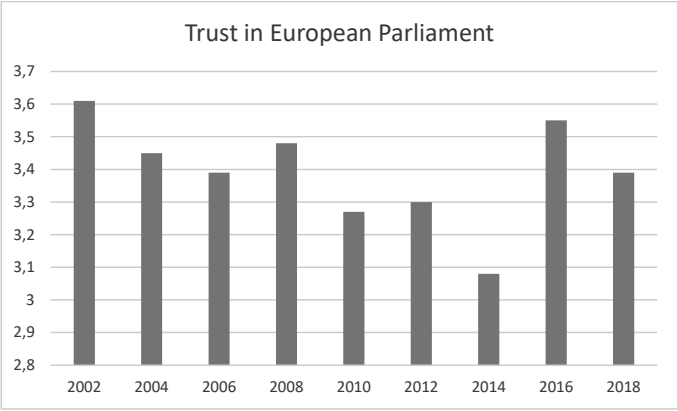
Regressionskoefficienterne β_j er vist Tabel 2.

Tabel 2. Parameter estimater

Parameter	*Estimate*	*Std error*	*t Value*
beta2	3.081	1.3259	2.32
beta3	2.363	1.0419	2.27
beta4	0.588	0.4460	1.32
beta5	-0.158	0.4377	-0.36
beta6	1.034	0.5914	1.75

Udviklingen i TRSTEP (=trust in the European Parliament)

Der er jo sket i England i den årrække, hvor analyserne blev foretaget. Indsamlingerne af data sker om efteråret. Så i 2016 er indsamlingen foretaget efter Brexit, da afstemningen foregik den 23. juni 2016.



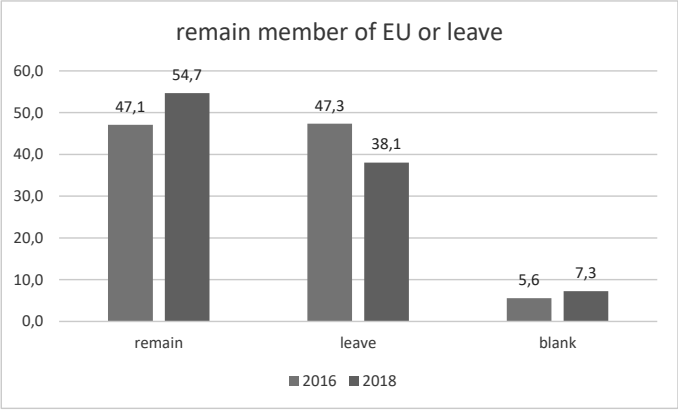
Man bemærker en nedadgående trend fra 2008 til 2014. Faktisk var "trusten" til Europarlamentet den laveste i år 2014. En voldsom stigning skete umiddelbart efter afstemningen om Brexit.

Efter Brexit afstemningen i 2016 blev der følgende spørgsmål også stillet:

"Would vote for XXX to remain member of European Union or leave" (dette spørgsmål er også stillet I England). Spørgsmålet blev stillet i alle EU's medlemslande, deraf de fire X-er. Skalaen var

Remain a member of the European Union	Leave the European Union	Would submit a blank ballot paper	Would spoil the ballot paper	Would not vote in EU referendum	Not eligible to vote

I figuren er vist den procentvise fordeling for dette spørgsmål.



Man kan notere sig, at der I 2018 var et klart flertal for at forblive I EU.

Can life events predict first-time suicide attempts?

A nationwide longitudinal study

Mogens Christoffersen, VIVE

Abstract

The prevention paradox describes circumstances in which the majority of cases with a suicide attempt come from a population of low or moderate-risk, and only a few from a 'high-risk' group. The assumption is that a low base rate in combination with multiple causes makes it impossible to identify a high-risk group with all suicide attempts.

Administrative registers were used to identify a group at higher risk of suicidal behaviour within a population of six national birth cohorts (N = 300,000).

Lifetime prevalence were 4.5% for first-time suicide attempts. Family background and family child-rearing factors were predicative of later first-time suicide attempts. A young person's diagnosis with psychiatric or neurodevelopmental disorders (ADHD, anxiety, depression, PTSD), and being a victim of violence or sex offences contributed to the explanatory model. Contrary to the prevention paradox, results suggest that it is possible to identify a discrete high-risk group (<12%) among the population from whom two thirds of all first-time suicide attempts occur, but one third of observed suicide attempts derived from low- to moderate-risk group.

Findings confirm the need for a combined strategy of universal, targeted and indicated prevention approaches in policy development and in strategic and practice responses, and some promising prevention strategies are presented.

Lack of progress in predicting suicide

Significant attention has been invested in identifying those most likely to die by suicide (Christiansen, Larsen, Agerbo, Bilenberg, & Stenager, 2013; Evans, Hawton, & Rodham, 2004; Glenn et al., 2018; Large, Sharma, Cannon, Ryan, & Nielssen, 2011; Maris, Berman, Maltsberger, & Yufit, 1992; Ribeiro et al., 2016).

However, little headway has been made in the areas of prediction, explanation, and prevention. Meta-analysis has found that predicative ability has not improved across 50 years of research (Franklin et al., 2017; Large et al., 2016; Velupillai et al., 2019). We lack sufficient understanding of what combined risk factors pre-dispose individuals to attempt suicide (Franklin et al., 2017; Glenn & Nock, 2014). This is particularly the case with young people (Cha et al., 2018) and with initial attempts as opposed to repeat suicide attempts (Glenn & Nock, 2014). Improving knowledge about risk factors for, and effective prevention strategies to reduce, first-time suicide attempts is critical, as we know that most of those who die by suicide do so following an initial attempt (Busch & Jacobs, 2003).

Lack of progress is partly due to (a) the low base rate of suicidal behaviour, and (b) the significant challenge of demonstrating scientifically that any single intervention has been effective in preventing suicide (Goldney, 2000). As yet, no single intervention has been shown in a well-documented randomised control trial to reduce suicide. Only a few studies have met randomised-controlled trial research criteria; even when results from similar trials are synthesised using meta-analytical techniques, numbers are insufficient to detect differences (Hawton et al., 1998; Mann et al., 2005; Zalsman et al., 2016).

The present study

Data and measures

To demonstrate even small changes in suicidal behaviour, huge samples are needed. A Danish register-based dataset followed all individuals born from 1980 up to and including 1985 in the present study. Six birth-cohorts are followed from age 15 to 29 years.

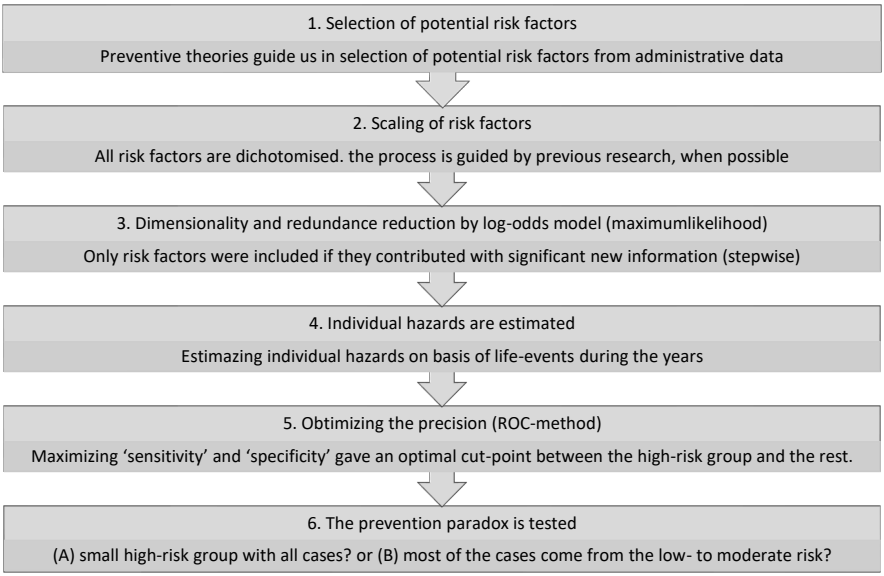
A stand-alone national register of suicide attempts has not yet been established, but in 1996 a methodology was constructed to combine diverse variables stored in national administrative registers. The definition of first-time suicide attempts includes behaviour that conforms to the following three conditions: (i) suicide attempts that lead to hospitalisation, (ii) assessment, of any injury or trauma as an act of self-mutilation based on the international statistical classification of injuries when discharged from hospital, (iii) damage included in a specified list of traumas traditionally connected with suicide attempts e.g. cutting of the wrist (carpus), firearm wounds, hanging, self-poisoning with drugs, ingestion of pesticide, cleaning fluids, alcohol or exposure to carbon monoxide. Intentional self-harm, based on hospital admissions to a psychiatric ward, is also included (Christoffersen, M. N., Poulsen, & Nielsen, 2003). This method has been used to analyse suicidal behaviour in several Scandinavian studies (Christiansen et al., 2013; Helweg-Larsen et al., 2006; Jablonska et al., 2009).

Analysis

An event history from birth to adulthood is used to estimate hazards between 15 to 29 years. The inclusion of event history information in the statistical model refines the following research question: Can we determine whether a certain combination of events and other variables increases the risk of a first-time suicide attempt in subsequent years up to age 30?

Risk factors are sorted according to predictive value and capacity for explaining suicidal behaviour. We then test whether risk factors are precursors for first-time suicide attempts in this nationwide population study.

Figure 0. Data processing and hypotheses testing.



Note: 1. Preventive theories guide us in selection of potential risk factors. 2. All risk factors are standardized through construction of dummy-variables. The choice of cut-point is guided by previous research when possible. 3. The number of risk factors are reduced by a stepwise procedure. Only risk factors were included if they contributed with significant new information. 4. The individual hazards is estimated on the basis of the log-odds model for life-events during the formative years. 5. Maximizing both 'sensitivity' and 'specificity' gave an optimal cut-point between the high-risk group and the rest. 6. The prevention paradox is tested based on the individual hazards and the actual observations.

Statistical model

The best way to study events such as first-time suicide attempts and their causes is to collect event history data.

In order to boost predictive efficiency, the estimation of risk parameters is based on the maximum likelihood method via the discrete-time logistic odds-ratio regression model. It has been shown that the maximum likelihood estimator can be obtained by treating all the time units for all individuals as though they were independent, when studying events. The model is used to allow for changing covariates over time (Allison, 1982). Only first reported suicide attempts are analyzed. Individuals' event histories are broken up into discrete time units (age) from 15 to 29 years, in which an event either did or did not occur. The event history uses age as the time unit (t = 15, 16, 17…29). The study only includes persons living in Denmark at age 15 which represents 300,000 individuals and 4 million person-years. A dummy variable was included in the model for each age from 15 to 29 in order to adjust for age. Each individual is observed until such a point where an event either occurs, or the observation is censored by reaching the age limit, because of death, or because the individual is lost to observation for other reasons. Individuals were excluded after the first event.

Risk factors were all selected from previously described prevention theories, and successively included in the model, starting with the most significant risk factors. Then only risk factors contributing a further predicative value were included. The stepwise procedure was carried out to select significant risk factors to give the best possible prediction, and avoid redundancies.

Some variables, such as gender, may be constant over time, while others, such as living in a disadvantaged area, depression, anxiety, PTSD, may vary. Potential risk factors were categorised into three types. Type 1 covers risk factors assumed to cover the years before and after the years observed (e.g. parental mental illness, a child's somatic disadvantages). Type 2 covers risk factors observed at age t and is assumed only to be present at t (e.g. living in a disadvantaged area, child's diagnose with anxiety, depression). Type 3 covers risk factors observed at time t and is assumed also to be present all the following years (e.g. victim of violence, family separation).

Measuring predictive accuracy

A key-question is how to assess the accuracy of predictions of first time suicide attempts. Various criteria have been developed in order to answer this question.

We contrast what is expected, according to the event history of every single individual person, with what is actually observed. We want to maximise both 'sensitivity' and 'specificity' which means the number of expected first-time suicide attempts in the model which convert into actual events. We also want to maximise 'specificity,' which is the number of individuals neither attempting suicide according to observations, nor according to expectations from the model. Special analytical tools (Receiver Operating Characteristic, ROC or Relative Improvement over Chance, RIOC) were developed to maximise both sensitivity and specificity by varying the cut-off point between high-risk groups and low- or moderate risk-groups. The ROC method is unaffected by base rates (Loeber & Dishion, 1983; Mossman, 1994; Mossman, 2000).

Finally, we estimate the number of observed first-time suicide attempters in the high-risk group and compare findings with observed numbers derived from the low- to moderate risk-group thus testing Rose's prevention paradox.

Administrative register data

Danish administrative registers used in this study include: Population Statistics, Medical Register on Vital Statistics, Causes of Death Register, Population and Housing Census, Unemployment Statistics, Education Statistics, Social Assistance Act Statistics, Income Compensation Benefits, Labour Market Research, Fertility Research, Criminal Statistic Register, National Patient Register, Danish Psychiatric Nationwide Case Register and Medical Birth Register. Personal identity numbers are used, at initial stages, to link information for each individual from birth, together with information about their parents. Later, personal identity numbers are substituted with an encrypted number for security and ethical reasons. Researchers only have access to these encrypted numbers as part of the analysis.

Data are not aggregated or clustered but individually collected, independently registered prospectively; and collected independently from several agencies. Data completely cover all calendar years from birth to early adulthood.

If we succeed in identifying a small high-risk group (less than 12%) who are most likely to represent all cases of first-time suicide attempts, we will have a more precise picture of the

history of events preceding such activity, which will bring us a step closer to identifying key factors associated with these events.

Results

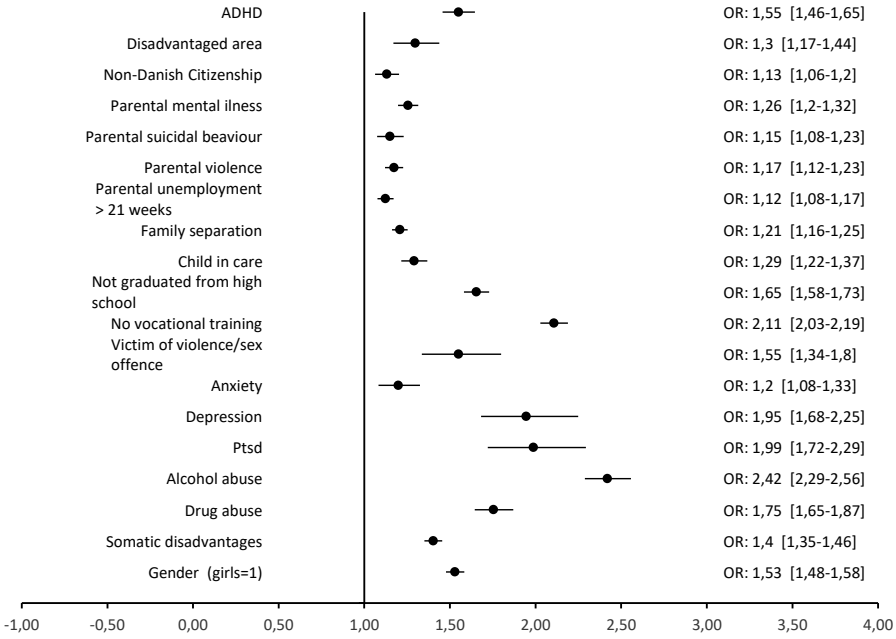
Overall, this study shows a low base rate of suicide attempts. Of the 300,000 individuals born between 1980 and 1985, and followed from age 15 to their 29 years. 13,358 persons attempted suicide at least once during the 15-year period of study. Lifetime prevalence was approximately 4.5 percent.

Nineteen risk factors, from the previously outlined theoretical risk groupings, were linked with increased risk of a first-time suicide (Figure 1). We calculated the adjusted discrete time log-odds -model, the covariates, the percentage of person-years exposed to respective risk factors, the estimates of covariates, standard errors and odds ratios. Nine of the nineteen risk factors had an effect size of more than 1.5 (odds ratio). Four of these had an effect size of just under or over 2, when adjusted for gender and age. All 19 risk factors are significant and add new information in terms of refining relevant risk factors for first-time suicide attempts. Adjusted effect sizes OR and statistical uncertainties are seen in Figure 1.

Developmental adversity and exposure to disadvantage during formative years, as well as decision-making processes linked to membership of a high-risk group appeared the most influential risk factors. Certain high-risk groups emerged with significantly higher risk of first-time suicide. These included young people (unadjusted effect sizes uOR): the risk of suicide attempts were OR: 2.11 (uOR: 3.45) with no vocational training, OR: 1.65 (uOR: 2.84) who had not graduated from high school, OR: 1.40 (uOR: 1.91), with somatic disabilities OR: 2.42 (uOR: 7.83), with alcohol abuse, OR: 1.75 (uOR: 11.30) drug abuse. Victims of violent crime or of sex offences OR: 1.55 (uOR: 2.78) were among the most at-risk groups, together with adolescents with depression OR: 1.95 (uOR: 22.98), severe stress and adjustment disorders e.g. PTSD OR: 1.99 (uOR: 22.28).

When considering adversity experienced during formative years, the following risk factors emerged as presenting relatively high risks: parental mental illness OR: 1.26 (uOR 2.12), parental suicidal behaviour OR: 1.15 (uOR: 2.28), parental violence OR: 1.17 (uOR: 2.03), family separation OR: 1.21 (uOR: 1.89), parental long-term unemployment OR: 1.12 (uOR: 1.67), placement in care OR: 1.29 (uOR: 4.07).

Figure 1. Forrest plot of subgroups, first-time suicide attempts, adjusted Odds Ratio (OR).



Parental alcohol abuse, according to hospital admissions, was predictive of an offspring's later suicide attempts (uOR: 2.05), but when adjusted for the other risk factors, parental alcohol abuse turned out to be non-significant. Being a teenage mother (uOR: 1.76) was predictive of an offspring's suicide attempt, but this was likewise non-significant in the adjusted regression model.

Theories of local community, social control, values and norms included indicators such as living in a socio-economically disadvantaged area. The number of years living in these areas seemed to be associated with an increase in risk of attempted suicide OR: 1.30 (uOR: 1.94). Being a non-Danish citizen was also associated with relatively higher risk of suicidal behaviour OR: 1.13 (uOR: 1.36).

In this study, ADHD was used as a proxy for risk-taking and impulsivity. Those presenting with a diagnosis of ADHD presented with relatively higher risk of suicidal behaviour OR: 1.55 (uOR: 9.97).

We found that about 57 percent of the population have a probability level of suicide less than 0.001, 12 percent have a probability level higher than 0.003 and 6 percent have a probability level higher than 0.004. In order to distinguish between high-risk groups and moderate or low-

risk groups we use a variable cut-off point. We adopted the ROC procedure (Receiver Operating Characteristic, ROC or Relative Improvement over Chance, RIOC) to resolve the dilemma of which cut-off point should be selected to provide the most accurate chance of preventing first-time suicide attempts.

The true positive rate (*sensitivity*) is defined as: a/(a+d), where 'a' is observed and expected cases and 'd' is observed but not expected. While *specificity* is defined as: b/(b+c), where 'b' is the not observed and not expected cases and 'c' is not observed, but expected cases (Woodward, 1999).

Sensitivity and specificity has been calculated and the maximum of the sum-curve is at the cut-off point 0.003, where the people most at risk of first time suicide attempts represent approximately 12 percent of the population. Unfortunately, the number of observed attempts in the high-risk group only counted 12 percent while 88 percent turned out to be false positives.

One third (31 percent) of attempted suicides that actually occurred would not be included in the treatment group if we used the most optimal cut-off point with a probability level of 0.003. Many young people would go under the radar, not receiving support and treatment.

Overall, we were not able to identify a high-risk group (less than 12 percent in the population) accounting for all first-time suicide attempts. On the other hand, two thirds of observed suicide attempts did occur in the high-risk group.

Conclusion, implications and limitations

Universal and selective prevention approaches are not mutually exclusive, but each tends to see risk through a different lens and prompts different responses to reducing suicide.

The current study sought to improve knowledge on suicide prevention through considering the extent to which Rose's Prevention Paradox is applicable and relevant to strategic planning in this field. It highlighted how decisions concerning the most effective approach to adopt, should be influenced by the extent to which prevention efforts can be targeted towards a clear-cut high-risk group to reduce risk effectively or, alternatively, whether risk is more dispersed across the broader population requiring more universal activity.

We found that life prevalence of first-time suicide attempts was 4.5 percent in the age from 15 to 29. A certain combination of 19 risk variables and events increased the risk of attempted suicide in the subsequent years up to age 30.

Strongest associations with suicidal behaviour were noted for young people: a) engaged in high-risk behaviours such as alcohol or drug reliance. b) who appeared to be not thriving in education/who lacked vocational training c) or who were victims of violence. The next strongest set of associations were for young people with d) risks relating to mental health conditions (PTSD, depression) or to ADHD and higher risk of suicide.

Generally, between 15 to 29 years, associations were marginally stronger for contemporary youth-based risks (e.g educational under-performance, mental health presentations) compared to more distal child-development-based risk factors or to risks linked to socio economic disadvantage. However, distal and socio economic risk factors still remained statistically significant in their influence on suicidal behaviour even during adolescent and young adult years. Furthermore, we know from other birth cohort studies, that many early socio demographic risks and child-development-based risk factors contribute to the likelihood of children experiencing later educational crises (Paget et al., 2018), engagement in risk-taking behaviours and experiencing

poorer mental health (Gutman, Joshi, Khan, & Schoon, 2018). This is particularly the case when exposure to risk occurs early and persists over time (Gutman et al., 2018; McLoyd, 1998; Smith, Brooks-Gunn, & Klebanov, 1997) .

Probability modelling, comparing projected deaths (based on the presence of risks outlined in this study) with actual deaths by suicide in Danish birth cohort studies, highlighted limitations in adopting an approach which simply focused suicide-prevention efforts on what appeared to be a high-risk group (<12%).

Some of the challenges highlighted by this study included: a) a large number of false positives (those with risk factors for suicide who do not go on to attempt suicide) occurring in this high-risk group if the net of high-risk individuals was too large; but, b) if the net of high-risk individuals was narrowed to increase accuracy, probability modelling noted a large number of false-negative cases sitting outside this high-risk target group (e.g. those who appear to have low risk but who subsequently go on to attempt suicide). This would result in preventative efforts missing many young people engaging in first time suicide attempts.

Findings did not support the prevention paradox since two thirds of first-time suicide attempts came from the identified high-risk group and suggesting that efforts directed toward this group could be helpful. Results also suggested a significant proportion of those at actual risk of suicidal behaviour sitting outside projected high-risk groups. Findings confirmed previous conclusions that any suicide-prevention strategy in high-income countries requires a multi-component portfolio of cross-sector effort and investment focused on universal, selective, and also indicated activity to prevent suicide (Van Der Feltz-cornelis, Christina M et al., 2011).

In terms of what such a multi-component cross sector strategy might look like, there remains a need for ongoing robust studies in this area to further our knowledge on effective interventions - and particularly to test out synergistic effects of multilevel preventive interventions (Van Der Feltz-cornelis, Christina M et al., 2011).

Limitations of this study

This study is based on a huge sample, with comprehensive information about potential risk factors and helps disentangle predictors influencing risk of first-time suicide. Still, unknown potential risk factors can be the Achilles' heel of any such strategy.

Bullying/peer victimisation have been strongly associated with higher risk of later suicide attempts (Geoffroy et al., 2016; Kim, Leventhal, Koh, & Boyce, 2009). This is one of the risk factors that this study was not specifically able to track via available administrative registers.

This study did not draw information from Danish registers on those presenting with early onset conduct problems - one of the most common mental health presentations affecting children. When these start early (affecting 5% of children) they are associated with a threefold increase in risk of suicide by age 27 years (Fergusson et al., 2006). Many children with ADHD also present with comorbid conduct problems and ADHD was included as a risk factor, but any future probability modelling should assess independent effects of childhood conduct disorder.

Effective social support might divert high-risk groups from suicide but we have not been able to find relevant indicators for social support in available administrative registers. Individual-level prevention strategies for those at higher risk of suicide include strengthening social support and coping skills; treatment and effective clinical care for underlying psychiatric and substance misuse disorders; development of problem-solving, cognitive behavioural and anger

management skills. (Hawton & James, 2005; Murphy, 1992; Murphy, Wetzel, Robins, & McEvoy, 1992; Public, 2001).

Activity targeting high-risk groups, restricting access to clothing and other materials that might permit hanging in psychiatric wards or prisons, could also be seen as a part of a selective intervention strategy but such data are missing in available administrative data.

The World Health Organisation's global comparative analysis of suicide (2014) highlighted commonalities but also distinct differences between high income and low to moderate income countries in relation to suicide. In this WHO analysis, Denmark was identified both as a high income country and as having experienced around a 27% reduction in overall suicide rates between 2000 and 2012. When considering relative risks, faced by those aged 15 to 29 years in Denmark (compared to those in other high-income countries), Denmark's young people appear to face low to moderate relative risk of suicide compared to peers in other high-income countries. Findings and recommendations from this study should therefore be considered in the context of this information.

Other relevant risks, such as death of a child, accidents, onsets of critical illnesses such as cancer, or being incarcerated have not been included in this current model, although they are closely linked to suicide risk with a potentially high effect size (OR). Problematically, predictive ability of rare events depends on two qualities: one is the number of people in the population who actually experience an event and the other is the odds ratio of suicide attempts among these individuals. Such traumatic events are rare and some may not find their way into administrative registers. Instead, we found a proxy for a traumatic event: a diagnosis of PTSD or anxiety during hospital admissions, and these proxies turn out to be effective predictors of first suicide attempts among young people.

Although we searched administrative databases for potential risk factors reviewing prevention theories for suicidal behavior, it can still be difficult to know what a specific risk factor captures. Results showed, for example, that alcohol abuse is one of the strongest correlates; it can be interpreted as part of a pattern of self-medicating for anxiety or depression, but it could also be indicative of an orientation toward risk. Without knowing what it captures, we are unsure about substance abuse's etiological role and what more targeted prevention might accomplish. Our methodology impels us to be cautious about cause-and-effect relationships and therefore also about prevention measures.

Administrative databases provide reasonably good prediction of first-time suicide attempts suggesting that it is possible to identify a discrete high-risk group (<12%) of the population from whom 69% of all first-time suicide attempts occur. From total numbers of first-time suicide attempts (n=13,358), two-thirds (9,181) came from the located high-risk group. Prediction quality is far from ideal and many potential risk factors are missing from databases.

The study provides some new information on the negative impact of multiple adversities and cumulative exposure of adverse childhood experiences. Future research could consider the impact of clustering risks on suicide risk and further analysis of moderators or mutually reinforcing family or environmental risk factors.

These findings confirm the need for a combined strategy of universal, targeted and indicated prevention approaches in policy development and in strategic and practice responses. Examples of such suicide-prevention interventions are summarised below.

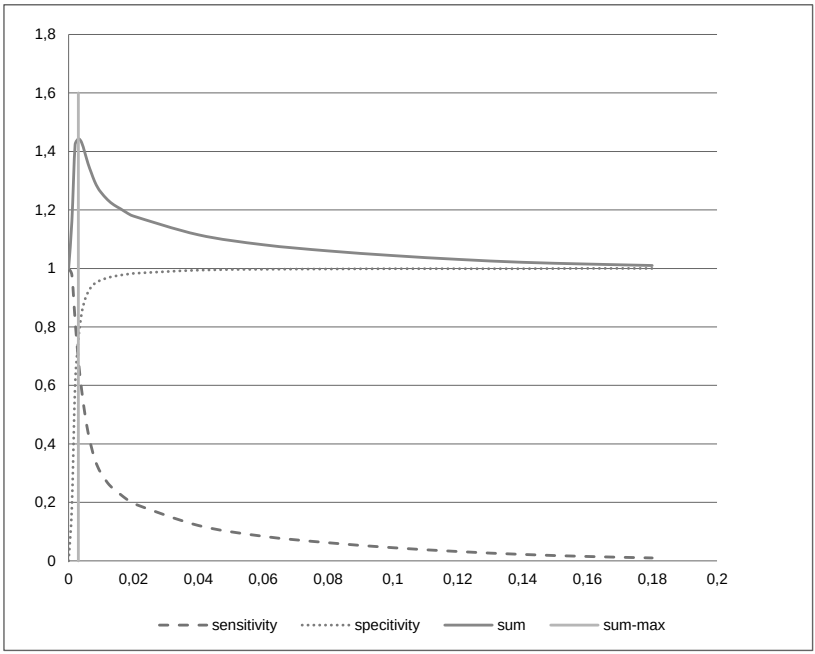
Appendix A. The outcome, risk factors and their definitions

Covariates and outcome variable:	***Definition***
Suicide attempt	Self-inflicted harm according to hospitals admissions. The definition of suicide attempts also included behaviour that conformed to the following conditions: (i) Suicide attempts that had led to hospitalization, (ii) assessment of the trauma being an act of self-mutilation according to the international statistical classification of injuries when discharged from hospital, (iii) the trauma had to be included in a specified list of traumas traditionally connected with suicide attempts: cutting in wrist (carpus), firearm wounds, hanging, self-poisoning with drugs, pesticide, cleaning fluids, alcohol or carbon monoxide. (We intend to exclude include non-suicidal self-harm, NSSH)
	Contemporary situation and opportunities
ADHD	Diagnosed with ADHD in a psychiatric ward according to the Danish Psychiatric Nationwide Case Register. ADHD F90 Hyperkinetic disorders and/or receiving ADHD-drugs 'N06BA04' or 'N06BA09'
	Local community, social control, values and norms
Disadvantaged Area	A governmental board has pointed identified the most disadvantaged housing areas. These are a part of the subsidized housing sector, consisting of 135 areas. About 4% of the population (200,000 persons) lives in these areas. Each area has 1,500 inhabitants, on average, ranging from. 30 to 14,000 persons (Graversen et al., 1997; Hummelgaard et al., 1997; Boligministeriet, 1993). These disadvantaged housing areas were divided into quintiles and the two most disadvantaged quintiles were identified as disadvantaged areas in the present by this dichotomized variable. These most disadvantaged areas would thus cover about 80,000 inhabitants or 1.6% of the total population.
Non-Danish Citizenship	The definition is based on fulfilling one of the following conditions: a) If at least one of the parents have Danish citizenship and is born in Denmark. b) If there is no information in the registers about any of the parents and the child himself/herself has Danish citizenship and is born in Denmark. All others are defined as non-Danish.
	Developmental theories, disadvantages during the formative years
Parental inpatient mental illness	One or both parents admitted to a psychiatric ward according to the Danish Psychiatric Nationwide Case Register
Parental suicidal behavior	Parents' suicide attempts according to the National Patient Register and the Danish Psychiatric Nationwide Case Register, or suicide according to the Causes of Death Register. Intentional self-harm according to hospitals admissions is also included.
Parental violence	Battered adults according to hospitals admissions. Parent exposed to assault or injuries of undetermined intent. Victims of violence which led to hospitalization and professional assessment that the injury was willfully inflicted by other persons. *Parent convicted for violence:* The Criminal Statistic Register records persons convicted for violence. This category comprises a wide range of criminal behaviour of various degrees of seriousness: manslaughter, grievous bodily harm, violence, coercion and threats. This category does not include accidental manslaughter in combination with traffic accidents, or rape, which belongs to the category of sexual offences.
Parental unemployed >21 weeks	Unemployment for at least one parent: The number of days unemployed (more than 21 weeks) during a calendar year. From registers of Income Compensation Benefits, Labour Market Research, and Unemployment Statistics. Parental unemployment for one or both parents.
Family separation	Information on all children who had experienced divorce, separation and or the death of a parent before they were 18 years old, taken from the Danish Central Population Register (CPR) that connects children to their parents whether they are married or not.
Child in care	The child is in care via home placement according to the children's acts section or the child is not living with the parents but in an institution or foster home according to population-based register of social assistance for children in care.
	Decision-making processes in high risk groups
Not graduated from high school	Passed basic, but had not gone on from school to university, not at least graduated, or ever been in high school (gymnasium)
Vocational qualification	Whether the person has a vocational or professional training (e.g. bricklayer, carpenter, dentist, lawyer, or teacher in a kinder garden). This does not include semi-skilled workers. Information is based on Education Statistics or the educational classification database which is population-based, including schooling and educational training for the highest education achieved by the person each parent in focus.
Victim of a violent or sexual offence	Persons who have been victim of violent personal crimes under the Danish Penalty Code (Jensen et al., 2003). Only criminal offences where the victim and the perpetrator are confronted e.g. offences against personal liberty, violence against the person, homicide, assault, robbery, threats, but not theft. Violation of an order of the Court or the police. Only incidents reported in the years 2001-2012. Persons who have been victim of a sexual crime under the Danish Penalty Code (Jensen et al., 2003). Only criminal offences where the victim and the perpetrator are confronted e.g. rape, sexual abuse, sexual exploitation, and indecent exposure. Violation of an order of the Court or the police. Only incidents reported in the years 2001-2012.
Anxiety	Diagnosed with Anxiety, panic disorder, generalized anxiety disorder, mixed anxiety and depression disorder, other mixed anxiety disorders and other anxiety disorder according to the Danish Psychiatric Nationwide Case Register. F41.0-9.
Depression	Diagnosed with mild, moderate or severe depression episode, single episode of depressive reaction, psychogenic depression or reactive depression according to the Danish Psychiatric Nationwide Case Register. F32.0-9.

PTSD	Diagnosed with reaction to severe stress, and adjustment disorders, acute stress reaction, post- traumatic stress disorder, adjustment disorders, or other reaction to severe stress according to the Danish Psychiatric Nationwide Case Register. F43.0-9.
Alcohol abuse	According to hospital admissions the following diagnoses were expected to be associated with long-term alcohol abuse: Alcoholic psychosis, alcoholism, oesophageal varices, cirrhosis of liver (alcoholic), chronic pancreatitis (alcoholic), delirium, accidental poisoning by alcohol. Mental and behavior disorder due to use of alcohol also included
Drug abuse	Addiction or poisoning by drugs according to hospitals admissions. Mental and behavioural disorder due to use of drugs (e.g. opioids, cannabinoids, cocaine). Dependence on morphine was not included if associated with diseases of chronic pain
Somatic disadvantages	Adolescents and young adults who had been hospitalized within the observation period for a severe handicap or chronic disease, other than mental handicap and psychiatric disease. Diagnoses included severe diseases of a chronic nature from all organ systems. Examples could be cancers, inborn errors, birth defects, cerebral palsy, and long lasting damages after head injuries necessitating hospitalization, epilepsy and sequelae after meningitis.

Parental PTSD	*Post-Traumatic Stress Disorder ICD-10: F43,* One or both parents diagnosed in a psychiatric ward according to the Danish Psychiatric Nationwide Case Register
Parental anxiety disorders	*Anxiety disorders ICD-10: F40-F42.* One or both parents diagnosed with anxiety disorder in a psychiatric ward according to the Danish Psychiatric Nationwide Case Register, diagnosed in a psychiatric ward or receiving drugs treatment with one of the following psychopharmaca: N05BA12 Alprazolam, N0 6AB03 Fluoxetin, N06AB05 Paroxetin, N06AB06 Sertralin, N06AB10 Escitalopram, N06AX11 Mirtazapin, N06AX16 Venlafaxin, N03AE01 Clonazepam, N03AX16 Pregabalin. (Arendt et al., 2016; NICE, 2013)
Parental substance abuse	Alcohol abuse or drug abuse (see below)
Parental alcohol abuse	According to hospital admissions the following diagnoses were expected to be associated with long-term alcohol abuse: Alcoholic psychosis, alcoholism, esophageal varices, cirrhosis of liver (alcoholic), chronic pancreatitis (alcoholic), delirium, accidental poisoning by alcohol. Mental and behavior disorder due to use of alcohol also included (Christoffersen, Mogens Nygaard & Soothill, 2003; Christoffersen, Mogens Nygaard, 2016).
Parental drug abuse	Addiction or poisoning by drugs according to hospitals admissions. Mental and behavioral disorder due to use of drugs (e.g. opioids, cannabinoids, cocaine). Dependence on morphine was not included if associated with diseases of chronic pain
Parent diagnosed with ADHD	Diagnosed with ADHD in a psychiatric ward according to the Danish Psychiatric Nationwide Case Register.
Parental violence	*Battered adults according to hospitals admissions.* Parent exposed to assault or injuries of undetermined intent. Victims of violence, which led to hospitalization and professional assessment that other persons willfully inflicted the injury. *Parent convicted for violence:* The Criminal Statistic Register records persons convicted for violence. This category comprises a wide range of criminal behavior of various degrees of seriousness: manslaughter, grievous bodily harm, violence, coercion and threats. This category does not include accidental manslaughter in combination with traffic accidents, or rape, which belongs to the category of sexual offences.

***Figure 2. Sensitivity, specificity, and their sum against cut-points*.**



Note: Sensitivity: number of observed and expected in relation to number of observed. Specificity: number of observed and expected person, who did not attempt suicide in relation to observed individuals not attempting suicide. The vertical line (sum-max) represents a cut-point of 0.003.

GENERALIZED INFORMATION CRITERIA FOR SPARSE STATISTICAL JUMP MODELS

FEDERICO P. CORTESE, PETTER N. KOLM AND ERIK LINDSTRÖM

Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy
federico.cortese@unimib.it

Courant Institute of Mathematical Sciences, New York, NY 10012, USA
petter.kolm@nyu.edu

Centre for Mathematical Sciences, Lund University, Sweden
erik.lindstrom@matstat.lu.se

Abstract. We extend the generalized information criteria for high-dimensional penalized models to sparse statistical jump models, a new class of statistically robust and computationally efficient alternatives to hidden Markov models. In a simulation study, we demonstrate that the new generalized information criteria selects the correct hyperparameters with high probability. Finally, providing an empirical application, we infer the key features that drive the return dynamics of the largest cryptocurrencies. We find that a four-state model best describes the dynamics of cryptocurrency returns. The states have natural market-based interpretations as they correspond to bull, bull-neutral, bear-neutral, and bear market regimes, respectively.

1. Introduction

Simple statistical models such as linear regression are straightforward to fit to data. This is not the case for many of its more capable extensions, such as penalized splines or LASSO, that require hyperparameters to be specified (Ruppert et al., 2003; Tibshirani, 1996). Nystrup et al. (2020b) report substantial improvements by performing hyperparameter tuning of their statistical model within the specific context of its application, i.e. by choosing hyperparameters that optimize an application specific performance criteria. Novel advances in machine learning have generated greater attention to hyperparameter tuning, e.g. for the topology of deep neural networks (Hutter et al., 2019).

Frequently, the optimal values of the hyperparameters are found using *information criteria* (IC) such as the *Akaike's information criterion* (AIC) or the *Bayesian information criterion* (BIC)(Akaike, 1974; Schwarz, 1978). Recently, Fan et al., 2013 extended the usage of IC

Date: December 9, 2022.
Key words and phrases. Clustering; Cryptocurrencies; Feature Selection; Information Criteria; Model Selection; Regime Switching; Unsupervised Learning.

to situations where the number of features, P, are of the same magnitude or substantially larger than the number of observations, T.

Regime switching models such as hidden Markov models (HMMs) are frequently applied when the data exhibits non-stationarities such as those occurring in financial time series (Zucchini et al., 2017). In particular, HMMs are capable of describing the majority of stylized facts documented in financial markets (Rydén et al., 1998). However, it can be challenging to use them in practice as their log-likelihood functions are non-convex, and their estimation is sensitive to model misspecification and parameter initialization (Rydén, 2008).

The class of so-called *statistical jump models* (JMs) introduced in Bemporad et al. (2018) provides an interesting alternative to HMMs, as they implement a framework that generates a complex model by switching between simpler ones. The same authors show that a simple HMM is a special case of their framework. Nystrup et al. (2020c) and Nystrup et al. (2020a) build upon this work, replacing the local models with K-means clustering. The resulting algorithm fits a temporal clustering model, where the persistence of cluster membership is controlled by a hyperparameter. Nystrup et al. (2021) propose a framework to perform feature selection for temporal clustering referred to as *sparse statistical jump models* (SJMs). In contrast to classical HMMs, the algorithm used for fitting them converges rapidly even when considering a large number of irrelevant features and is remarkably robust against model misspecification and poor initialization. However, SJMs require proper tuning of several hyperparameters that relate to their temporal persistence, feature selection and the number of clusters used. While the tuning can be done by hand, it is highly desirable to develop techniques to perform it automatically.

The purpose of this article is to adapt the *generalized information criteria* (GIC) framework of Fan et al. (2013) to the class of SJMs. To construct suitable IC for hyperparameters selection for SJMs, we derive expressions for their model fit and complexity. To compare the resulting IC, we conduct a simulation study and find that a BIC suitably extended for SJMs consistently outperforms the alternative IC. Thereafter, we present an empirical study where we fit an SJM on a large set of features related to the cryptocurrency markets by selecting the best model based on the extended BIC.

2. Methodology

2.1. **Information Criteria.** The log-likelihood function contains much information about the data and the model but it cannot be used for comparisons between competing models since it increases with increasing model complexity. Therefore, hyperparameters are often selected using cross-validation or IC. The latter is defined as a combination of two terms

$$IC := F + a_T M, \tag{1}$$

where F is a measure of model fit, M is a measure of model complexity, and a_T a positive sequence depending on the sample size T and possibly, the number of parameters and/or features considered (see, for example, Konishi et al. (2008) and Fan et al. (2013)). It is common in classical statistics to approximate model complexity with the number of model parameters. However, the concept of model complexity is more general in modern statistics and machine learning, where besides model parameters in the traditional sense, the models frequently have hyperparameters and regularization terms to address the complexity of large sets of candidate features. Parameters refer to the set of *internal* quantitative attributes that defines the model itself, hyperparameters are *external* to the model as their values are chosen before model estimation, and features are constructed from the data. In the following, we refer to the number of model parameters as q and the number features as P. We denote by $\ell(\hat{\theta})$ the maximized value of the log-likelihood function for a model with q parameters. Then we note that the AIC and BIC, the two most common IC, defined as

$$\text{AIC} := -2\ell(\hat{\theta}) + 2q\,, \tag{2}$$

$$\text{BIC} := -2\ell(\hat{\theta}) + q\log(T)\,, \tag{3}$$

take the form of IC in equation (1) where $-2\ell(\hat{\theta})$ measures the model fit, q denotes the number of model parameters and serves as a measure of model complexity, and $a_T = 2$ (for AIC) and $a_T = \log(T)$ (for BIC), respectively. Frequently, the AIC is used to select the best predictive model and is an asymptotically unbiased estimator of the expected Kullback-Leibler (KL) risk loss under the assumption that the candidate model family includes the true model (Hurvich et al., 1990). Schwarz (1978) derives the BIC by considering the posterior probability of a specific model given data. It is known that this IC finds the correct model asymptotically.

Fan et al. (2013) consider a setup where the number of features grows substantially faster than the number of observations. They define a GIC as

$$\text{GIC} = \frac{1}{T}\left\{2\left(\ell_T(Y,Y) - \ell_T(\hat{\theta}_{\mathcal{A}}, Y)\right) + a_T M\right\}\,, \tag{4}$$

where $\ell_T(Y,Y)$ is the log-likelihood for the saturated model and $\ell_T(\hat{\theta}_{\mathcal{A}}, Y)$ is the log-likelihood using the estimated active set of parameters, $\mathcal{A}$. They suggest setting the model complexity equal to the cardinality of the active set of features (i.e. $M = |\mathcal{A}|$), and $a_T = \log(\log(T))\log(P)$. We will refer to this particular choice of GIC as the *Fan-Tang information criterion* (FTIC).

2.2. **Regime switching and sparse statistical jump models.** While HMMs have a long tradition in statistics they are known to have undesirable properties. For example, there are as many singularities in the log-likelihood surface as there are observations. We examine a larger class of models in this article, that of statistical jump models (not to be confused

with jump-diffusion models), introduced in Bemporad et al. (2018) as a framework for (a) combining several simpler models into more complex ones, and (b) generalizing a number of well-known model classes, including HMMs. In this article we focus on the SJM of Nystrup et al. (2020c). The core idea of their framework is regime classification in an HMM is closely related to temporal clustering in a sufficiently high-dimensional space that encodes the relevant statistical properties. Specifically, they minimize the loss function given by

$$\sum_{t=1}^{T-1} \left[\|\widetilde{\boldsymbol{y}}_{t,P} - \boldsymbol{\mu}_{s_t}\|_2^2 + \lambda \mathbb{I}_{s_t \neq s_{t-1}} \right] + \|\widetilde{\boldsymbol{y}}_{T,p} - \boldsymbol{\mu}_{s_T}\|_2^2 , \tag{5}$$

where $\widetilde{\boldsymbol{y}}_{t,P} \in \mathbb{R}^P$ is a P-dimensional standardized feature vector obtained from a given time series data set, $\{s_t\}$ is a K-dimensional latent state sequence, $\boldsymbol{\mu}_{s_t}$ is the conditional mean of state s_t, and the hyperparameter $\lambda \geq 0$ controls the frequency of jumps between states. Nystrup et al. (2020c) propose a coordinate descent approach to estimate the model parameters and state sequence by alternating between (i) minimizing a quadratic loss function, and (ii) solving a dynamic programming problem with Viterbi-like algorithm (Viterbi, 1967). The resulting algorithm has many attractive properties like robustness against poor initialization, distributional or sojourn time misspecification and fast convergence.

Nystrup et al. (2021) further extend their algorithm by incorporating feature selection, allowing for a large number of features to influence the inferred state sequence, a property that has been difficult to incorporate in traditional HMMs. Their work is based on the Witten et al. (2010) framework. They observe that the *total sum of squares* (TSS) can be expressed as the sum of *within-cluster sum of squares* (WCSS) and the *between cluster sum of squares* (BCSS), that is

$$\underbrace{\sum_{t=1}^{T} \|\widetilde{\boldsymbol{y}}_{t,P} - \bar{\boldsymbol{\mu}}\|_2^2}_{=:\text{TSS}} = \underbrace{\sum_{t=1}^{T} \|\widetilde{\boldsymbol{y}}_{t,P} - \boldsymbol{\mu}_{s_t}\|_2^2}_{=:\text{WCSS}} + \underbrace{\sum_{k=1}^{K} n_k \|\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}\|_2^2}_{=:\text{BCSS}} , \tag{6}$$

where n_k denotes the number of observations belonging to the k-th cluster, and $\bar{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}_k$ are the unconditional and conditional means of the features in the k-th state, respectively. Therefore, as minimizing the WCSS is equivalent to maximizing the BCSS, Nystrup et al., 2021 propose to solve

$$\begin{aligned} \max_{\boldsymbol{\mu}_k, \{s_t\}, \boldsymbol{w}} \quad & \boldsymbol{w}' \textstyle\sum_{k=1}^{K} n_k \left(\boldsymbol{\mu}_k - \overline{\boldsymbol{\mu}}\right)^2 - \lambda \sum_{t=1}^{T-1} \mathbb{I}_{s_t \neq s_{t+1}} \\ \text{subject to} \quad & \|\boldsymbol{w}\|^2 \leq 1, \quad \|\boldsymbol{w}\|_1 \leq \kappa \\ & w_p \geq 0 \quad \forall p , \end{aligned} \tag{7}$$

where $\boldsymbol{w} = (w_1, \ldots, w_P)'$ is the vector of feature weights and the hyperparameter $1 \leq \kappa \leq \sqrt{P}$ controls the degree of sparsity of the features. The resulting model is called a *sparse statistical jump* (SJ) *model*. While the resulting algorithm is more complex, it can still be solved by

coordinate descent. We refer the reader to the original article by Nystrup et al. (2021) for details and publicly available `python` code[1].

2.3. **GIC for sparse jump models.** In this section we extend the GIC to the SJM. First, we substitute the log-likelihood $\ell_T(\cdot,\cdot)$ with the WCSS (in equation (6)), and denote it by $L_T(\lambda,\kappa,K;\boldsymbol{Y})$ where λ, κ, K and $\boldsymbol{Y}$ are the values for the jump penalty, sparsity hyperparameter, number of latent states and the matrix of features, respectively.

An important question is how to choose the hyperparameters $\bar{\lambda}$, $\bar{\kappa}$ and $\bar{K}$ corresponding to the saturated model in the SJM setting. While it is straightforward to set $\bar{\lambda}=0$ and $\bar{\kappa}=\sqrt{P}$ (as this choice leads to an SJM with no jump penalty that considers the entire set of features), it is not clear how to choose $\bar{K}$. Based on our experience, if the goals is to estimate a model with recurrent states, we suggest to not exceed $\bar{K}=6$ for the examples we considered. Indeed, when $\bar{K}$ is too high the GIC selects a large number of states, each one being visited only once. This type of time series clustering falls into the change-point detection framework (Aue et al., 2013), which is not the topic of the present paper. For the SJM, we define the model complexity measure M by

$$M = K_0|\mathcal{A}_0| + |\mathcal{A}_0|(K-K_0) + K_0(|\mathcal{A}|-|\mathcal{A}_0|) + \sum_t \mathbb{I}_{s_t\neq s_{t-1}}\,. \tag{8}$$

The three first terms come from a linear approximation of $K|\mathcal{A}|$ near the point $(K_0,|\mathcal{A}_0|)$, and hence penalize for increasing values of K and $|\mathcal{A}|$, the number of latent states and active features. $|\mathcal{A}|$ depends indirectly on the hyperparameter κ; it increases with increasing values of κ. In practical applications, we recommend to choose K_0 and $|\mathcal{A}_0|$ based on prior knowledge of the number of latent states and relevant features. The last term in equation (8) counts the number of jumps across states and is binomially distributed, with the number of jumps being the sufficient statistic for a binomial distribution. This term depends indirectly on the jump penalty λ, as the number of states changes increases with decreasing values of λ. Altogether, the resulting GIC for the SJMs is given by

$$\text{GIC} = \frac{1}{T}\left\{2\left(L_T(\bar{\lambda},\bar{\kappa},\bar{K};\boldsymbol{Y}) - L_T(\lambda,\kappa,K;\boldsymbol{Y})\right) + a_T M\right\}\,. \tag{9}$$

From (9) it is straightforward to obtain new versions of AIC, BIC and FTIC for the SJM by setting the value of a_T equal to 2, $\log T$ and $\log(\log(T))\log(P)$.

3. Simulation Study

We compare the proposed AIC, BIC and FTIC for the SJMs in a simulation study, evaluating their ability to correctly identify hyperparameters values. For this purpose, we simulate a dataset $\boldsymbol{Y}\in\mathbb{R}^{T\times P}$ ($T=1000$ and $P=100$), from a Gaussian HMM for varying number of latent states $K_{\text{true}}=2,3,4$. We set the vector of initial probabilities $\boldsymbol{\pi}=\{1/K_{\text{true}}\}$,

[1] `https://doi.org/10.1016/j.eswa.2021.115558`

K_{true}	IC	Value	λ	κ	K	ARI($\{\hat{s}_t\}$)	ARI($\{\omega_p\}$)
2	AIC	3.41	0	10	4	0.15	0.03
	BIC	7.44	5	7	2	**1.00**	**0.66**
	FTIC	9.10	5	6	2	**1.00**	0.50
3	AIC	3.25	5	8	3	**1.00**	**0.79**
	BIC	5.38	5	8	3	**1.00**	**0.79**
	FTIC	7.40	10	7	3	0.95	0.65
4	AIC	1.55	5	9	3	**1.00**	**0.95**
	BIC	4.24	5	8	4	**1.00**	0.81
	FTIC	6.57	5	7	4	**1.00**	0.64

TABLE 1. Simulation results for the minimum AIC, BIC and FTIC and the corresponding λ, κ and K for varying number of true latent states K_{true}. Value refers to the estimated value of the reported IC, and ARI($\{\hat{s}_t\}$) and ARI($\{\omega_t\}$) are the ARIs computed between true and estimated sequences of states, and between true and estimated sequences of active features, respectively.

matrix of transition probabilities with $\pi_{ii} = 0.8$ $(i = 1, \ldots, K_{\text{true}})$ on the main diagonal, and $\pi_{ij} = (1-\pi_{ii})/(K_{\text{true}}-1)$ $(i, j = 1, \ldots, K_{\text{true}})$ elsewhere. We draw the state-conditional mean vectors $\boldsymbol{\mu}_p^{(k)}$ $(p = 1, \ldots, P,\ k = 1, \ldots, K_{\text{true}})$ from the uniform distribution $\mathcal{U}(-2, 2)$. The state-conditional covariance matrices have diagonal elements equal to one and off-diagonal elements given by $\boldsymbol{\rho} = \left(\rho_{ij}^{(1)}, \ldots, \rho_{ij}^{(K_{\text{true}})}\right)', \forall i, j = 1, \ldots, P, i \neq j$, given by: (a) $\boldsymbol{\rho} = (0.8, 0.4)'$ when $K_{\text{true}} = 2$; (b) $\boldsymbol{\rho} = (0.8, 0.6, 0.3)'$ when $K_{\text{true}} = 3$; and (c) $\boldsymbol{\rho} = (0.8, 0.6, 0.3, 0.0)'$ when $K_{\text{true}} = 4$. Similar to the simulation study of Nystrup et al. (2021), we consider the augmented dataset $\widetilde{\boldsymbol{Y}}$ consisting of $\widetilde{P} = 500$ standardized features, where the first 100 correspond to the standardized version of $\boldsymbol{Y}$ and the other 400 are constructed through a random permutation of its rows. Using this augmented dataset allows us to test the ability of the GIC to select the correct features and ignore the irrelevant ones.

We fit a family of SJMs by varying λ and κ such that $\lambda \in \{0, 5, 10, 25, 50, 100\}$ and $\kappa \in \{1, 2, 3, \ldots, \lfloor\sqrt{\widetilde{P}}\rfloor\}$, where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. We compute the three IC for each possible λ, κ, K according to formula (9). We set the hyperparameters for the saturated model to $\bar{\lambda} = 0$, $\bar{\kappa} = \sqrt{\widetilde{P}}$, and $\bar{K} = 6$. To determine M, we take K_0 and $|\mathcal{A}_0|$ to be equal to their true counterparts, i.e. $|\mathcal{A}_0| = 100$ and $K_0 = K_{\text{true}}$.

To assess the ability of the IC in selecting the correct values of λ and κ, we compute the adjusted Rand index (ARI) (Hubert et al., 1985) between true and estimated sequences of states $\{\hat{s}_t\}$, and between true and estimated sequences of active features $\{\omega_p\}$. We obtain $\{\omega_p\}$ through an indicator function which is equal to one when feature p is in $\mathcal{A}$ and zero otherwise. We denote the two resulting ARIs by ARI($\{\hat{s}_t\}$) and ARI($\{\omega_p\}$), respectively. Recall that an ARI equal to one corresponds to a perfect match between the elements of the two sequences, and the index decreases as similarity decreases.

Table 1 shows the AIC, BIC, FTIC and the two ARIs when the true number of latent states is equal to $K_{\text{true}} = 2, 3, 4$, respectively. The results suggests that BIC and FTIC are better suited to determine the jump penalty λ, whereas AIC and BIC appears to be better suited for detecting the value of κ. In fact, the minimum BIC and FTIC correspond to a value of λ that results in $\text{ARI}(\{s_t\}) = 1$ for most scenarios. In contrast, in most of the scenarios, the minimum AIC and BIC result in values of κ that maximize $\text{ARI}(\{\omega_p\})$.

For each $K_{\text{true}} = 2, 3, 4$, we fit SJMs by varying λ, κ and K and then calculate the resulting AIC, BIC and FTIC. Figure 1 depicts how the IC depend on the hyperparameters, where the red line with circles, the green with triangles and the blue with squares represents the IC value for $K = 2, 3, 4$, respectively. To illustrate how an IC depends on λ and κ, for each IC and K_{true} we generate two separate panels. In one panel the x-axis represents λ, and in the other κ. Interestingly, AIC is not able to select the true number of latent states when $K_{\text{true}} = 2$ and $K_{\text{true}} = 3$. In contrast, BIC and FTIC attain their minima when the number of latent states corresponds to K_{true} in all scenarios, with BIC providing somewhat better results in terms of the relative distance between IC computed for different K.

4. An Application to Cryptocurrencies

In this application we aim to determine what are the most important drivers of the return dynamics of the four largest and most liquid cryptocurrencies: Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC) and Bitcoin Cash (BCH). We obtain cryptocurrency prices and trading volumes from the Crypto Asset Lab, an independent lab established at the University of Milano-Bicocca. Prices are volume-weighted, recorded at midnight UTC, from the Coinbase-pro, Poloniex, Bitstamp, Gemini, Bittrex, Kraken, and Bitflyer digital exchanges. Similarly, the daily volume of each cryptocurrency is calculated as the sum of the individual volumes from the same exchanges. We also consider variables related to the network activity, obtained from intotheblock.com. From this data, we construct a large set features that are potential candidates for explaining cryptocurrency returns. In particular, we use first difference of the logarithm of prices and volume and then compute *exponential moving averages* (EMA) for log-returns and volatilities with half-lives equal to 1, 2, 7 and 14 days. We construct exponentially weighted linear and Gerber correlations (Gerber et al., 2022) between BTC log-returns and log-returns of all other cryptocurrencies and between log-returns and log-differences of trade volumes for each cryptocurrency, with the same half-lives as above. To capture network activity, for each cryptocurrency we use first differences of the aggregate volume of transactions recorded on-chain, log-differences of the total number of addresses with balance (AddWB), and log-differences of hash rates for BTC and ETH. In addition, we calculate exponentially weighted linear and Gerber correlations between each of the previously mentioned network metrics and the corresponding cryptocurrency log-returns, also

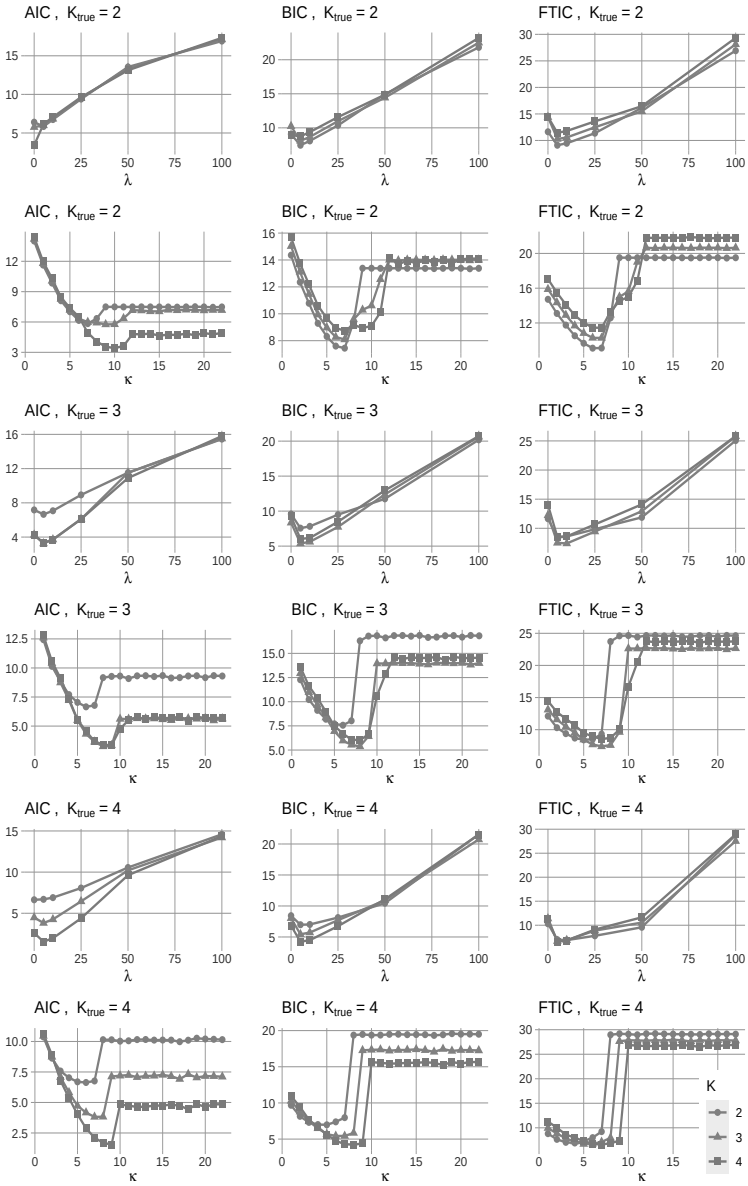


FIGURE 1. AIC, BIC and FTIC for different values of λ, κ and K. K_{true} denotes the true number of latent states. For each graph, the red line with circles, the green with triangles and the blue with squares represent the IC for $K = 2, 3, 4$, respectively.

with the same half-lives as before. Following Cong et al. (2021), we construct three value factors as the ratio between the total number of addresses and prices, the number of addresses with balance and prices, and the recorded volume on-chain and prices, for all four cryptocurrencies. Previous studies suggest that time series momentum is a crucial driver of cryptocurrency returns (Liu et al., 2021; Liu et al., 2022). Therefore, we add several momentum-based features such as the time series momentum signal of Moskowitz et al. (2012) with lags of 1, 2, 7 and 14 days, the relative strength index (RSI) and moving average convergence divergence signal indicators for all the cryptocurrencies. Finally, we consider the Amihud (2002) illiquidity measure, computed as the ratio of absolute daily log-returns and daily volumes for each coin. The final dataset consists of $P = 230$ features spanning the time period from January 1, 2019 through July 10, 2022, with $T = 1,352$ total number of daily observations. We fit a family of SJMs by varying λ, κ, K, and then select the best model based on the BIC, computed for $K_0 = 3$ and $|\mathcal{A}_0| = 100$. The BIC attains its minimum for $\lambda = 5$, $\kappa = 4$, $K = 4$. The resulting daily state-conditional averages of log-returns and volatilities of the four cryptocurrencies are 1.99%, 0.45%, -0.55%, and -2.99%, and 5.83%, 4.11%, 4.07% and 10.81%, respectively. Moreover, the average of the daily state-conditional pairwise correlations are 0.65, 0.77, 0.85, and 0.94. The sojourn times for each state are 20.18, 22.65, 14.69 and 9.38 days, and the SJM spends 41.8%, 38.5%, 14.1% and 5.5% of its time in each of them. These state-conditional statistics suggest an interpretation of the states as distinct market regimes, where the first state is a bull market regime, the second is a bull-neutral regime, the third is a bear-neutral regime and the fourth is a bear market regime.

Next, we examine the features selected by the SJM. Table 2 shows that out of the original 230, of which 20 are selected by the feature selection algorithm. EMAs have the largest weights, and their state-conditional values are coherent with the regime characterization above. Correlations between first differences of volumes and log-returns are decreasing from the first to the fourth regimes, indicating that higher market activity is associated with particularly turbulent phases (bull and bear). The RSIs, which have noticeably high weights, are above 70 in the bull regime, around 50 in the bull-neutral and bear-neutral regimes, and below 30 in the bear regime. We estimate a small weight (0.007) for the linear correlation between log-differences of the total number of addresses with balance and log-returns for BCH.

5. Conclusions

We extended the generalized information criteria for the purpose of performing hyperparameter selection for sparse statistical jump models. In a simulation study, we demonstrated that the new generalized information criteria infer the optimal values for the jump penalty

Feature	Weight	Bull	Bull-Neutral	Bear-Neutral	Bear
$\rho_{14}(\text{AddWB}_{\text{BCH}}, r_{\text{BCH}})$	0.007	0.181	0.034	-0.083	-0.149
$\rho_7(V_{\text{BTC}}, r_{\text{BTC}})$	0.029	0.340	0.155	-0.165	-0.420
$\rho_7(V_{\text{ETH}}, r_{\text{ETH}})$	0.037	0.385	0.114	-0.169	-0.380
$\rho_7(V_{\text{LTC}}, r_{\text{LTC}})$	0.024	0.434	0.134	-0.102	-0.234
$\rho_7(V_{\text{BCH}}, r_{\text{BCH}})$	0.001	0.455	0.162	-0.064	-0.219
$\rho_{14}(V_{\text{BTC}}, r_{\text{BTC}})$	0.032	0.284	0.116	-0.125	-0.317
$\rho_{14}(V_{\text{ETH}}, r_{\text{ETH}})$	0.043	0.307	0.075	-0.136	-0.280
$\rho_{14}(V_{\text{LTC}}, r_{\text{LTC}})$	0.030	0.365	0.114	-0.066	-0.164
$\text{EMA}_7(r_{\text{BTC}})$	0.062	0.012	0.005	-0.004	-0.017
$\text{EMA}_{14}(r_{\text{BTC}})$	0.071	0.010	0.004	-0.003	-0.012
$\text{EMA}_7(r_{\text{ETH}})$	0.075	0.016	0.007	-0.005	-0.022
$\text{EMA}_{14}(r_{\text{ETH}})$	0.071	0.012	0.005	-0.003	-0.013
$\text{EMA}_7(r_{\text{LTC}})$	0.076	0.016	0.004	-0.006	-0.023
$\text{EMA}_{14}(r_{\text{LTC}})$	0.091	0.013	0.004	-0.005	-0.015
$\text{EMA}_7(r_{\text{BCH}})$	0.074	0.018	0.003	-0.007	-0.025
$\text{EMA}_{14}(r_{\text{BCH}})$	0.088	0.013	0.003	-0.006	-0.016
RSI(BTC)	0.040	73.33	59.80	41.51	26.02
RSI(ETH)	0.060	73.77	60.73	41.27	27.51
RSI(LTC)	0.044	70.81	56.88	40.14	28.98
RSI(BCH)	0.045	70.93	55.16	39.04	26.80

TABLE 2. Estimated weights and state-conditional values of the selected features. RSI, ρ_d and EMA_d denote the relative strength index, exponentially weighted linear correlation and exponential moving average with a half-life of d days, respectively.

and sparsity hyperparameters, respectively, thereby obtaining the true number of latent states and features.

In an empirical application, we applied one of the new generalized information criteria to extract key features that drive the return dynamics of the largest cryptocurrencies. We found that the selected sparse jump model is consistent with economic intuition. In particular, the identified latent states represent bull, bull-neutral, bear-neutral and bear market regimes.

ACKNOWLEDGEMENTS

We acknowledge the University of Milano-Bicocca Crypto Asset Lab and Data Science Lab for supporting this work by providing data and computational resources. Petter N. Kolm and Erik Lindström were partly supported by the Knut and Alice Wallenberg Foundation under grant KAW 2020.0280.

REFERENCES

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Amihud, Yakov (2002). "Illiquidity and stock returns: Cross-section and time-series effects". In: *Journal of Financial Markets* 5.1, pp. 31–56.

Aue, Alexander and Lajos Horváth (2013). "Structural breaks in time series". In: *Journal of Time Series Analysis* 34.1, pp. 1–16.

Bemporad, Alberto, Valentina Breschi, Dario Piga, and Stephen P. Boyd (2018). "Fitting jump models". In: *Automatica* 96, pp. 11–21.

Cong, Lin William, George Andrew Karolyi, Ke Tang, and Weiyi Zhao (2021). "Value premium, network adoption, and factor pricing of crypto assets". In: *Network Adoption, and Factor Pricing of Crypto Assets (December 2021)*.

Fan, Yingying and Cheng Yong Tang (2013). "Tuning parameter selection in high dimensional penalized likelihood". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.3, pp. 531–552.

Gerber, Sander, Harry M. Markowitz, Philip A. Ernst, Yinsen Miao, Babak Javid, and Paul Sargen (2022). "The Gerber statistic: A robust co-movement measure for portfolio optimization". In: *The Journal of Portfolio Management* 48.3, pp. 87–102.

Hubert, Lawrence and Phipps Arabie (1985). "Comparing partitions". In: *Journal of Classification* 2, pp. 193–218.

Hurvich, Clifford M, Robert Shumway, and Chih-Ling Tsai (1990). "Improved estimators of Kullback–Leibler information for autoregressive model selection in small samples". In: *Biometrika* 77.4, pp. 709–719.

Hutter, Frank, Lars Kotthoff, and Joaquin Vanschoren (2019). *Automated machine learning: Methods, systems, challenges*. Springer Nature.

Konishi, Sadanori and Genshiro Kitagawa (2008). *Information criteria and statistical modeling*. Springer, New York, NY.

Liu, Yukun and Aleh Tsyvinski (2021). "Risks and returns of cryptocurrency". In: *The Review of Financial Studies* 34.6, pp. 2689–2727.

Liu, Yukun, Aleh Tsyvinski, and Xi Wu (2022). "Common risk factors in cryptocurrency". In: *The Journal of Finance* 77.2, pp. 1133–1177.

Moskowitz, Tobias J., Yao Hua Ooi, and Lasse Heje Pedersen (2012). "Time series momentum". In: *Journal of Financial Economics* 104.2, pp. 228–250.

Nystrup, Peter, Petter N. Kolm, and Erik Lindström (2020a). "Greedy online classification of persistent market states using realized intraday volatility features". In: *The Journal of Financial Data Science* 2.3, pp. 25–39.

– (2021). "Feature selection in jump models". In: *Expert Systems with Applications* 184, p. 115558.

Nystrup, Peter, Erik Lindström, and Henrik Madsen (2020b). "Hyperparameter optimization for portfolio selection". In: *The Journal of Financial Data Science* 2.3, pp. 40–54.

– (2020c). "Learning hidden Markov models with persistent states by penalizing jumps". In: *Expert Systems with Applications* 150, p. 113307.

Ruppert, David, Matt P Wand, and Raymond J Carroll (2003). *Semiparametric regression*. 12. Cambridge University Press.

Rydén, Tobias (2008). "EM versus Markov Chain Monte Carlo for estimation of hidden Markov models: A computational perspective". In: *Bayesian Analysis* 3.4, pp. 659–688.

Rydén, Tobias, Timo Teräsvirta, and Stefan Åsbrink (1998). "Stylized facts of daily return series and the hidden Markov model". In: *Journal of Applied Econometrics* 13.3, pp. 217–244.

Schwarz, Gideon (1978). "Estimating the dimension of a model". In: *The Annals of Statistics*, pp. 461–464.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the LASSO". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288.

Viterbi, Andrew (1967). "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory* 13, pp. 260–269.

Witten, Daniela M. and Robert Tibshirani (2010). "A framework for feature selection in clustering". In: *Journal of the American Statistical Association* 105.490, pp. 713–726.

Zucchini, Walter, Iain L. MacDonald, and Roland Langrock (2017). *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton, FL: CRC press.

Causality in Econometric Analyses

Arne Henningsen, Department of Food and Resource Economics (IFRO)

Most of research questions in economics are 'causal', i.e., how one variable affects another variable. In general, it is easiest to identify causal effects by running experiments but in economics, most research questions cannot be answered by experiments because experiments for answering these research questions would be unethical, excessively expensive, or for other reasons infeasible. Hence, most empirical research questions in economics can only be answered by using observational data, i.e., data generated by the 'real world' rather than by an experiment. Many different methods and identification strategies exist for estimating causal effects with observational data, e.g., matching methods, instrumental variable methods, difference-in-differences, synthetic control methods, regression discontinuity design, and `causal' machine learning methods. The presentation at the 'Symposium i Anvendt Statistik' will give an overview over methods and identification strategies that are used to estimate causal effects with observational data in economics and it will discuss the strengths and weaknesses of these methods.

Nyheder i SAS

Anders Milhøj

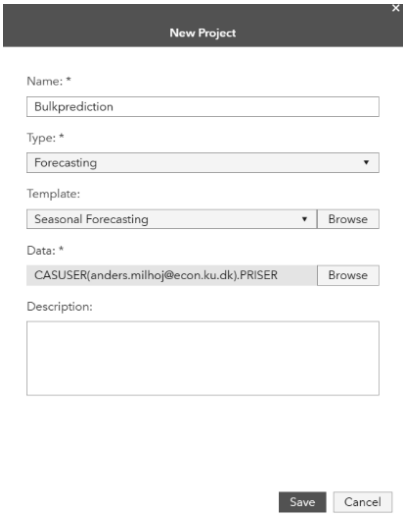
anders.milhoj@econ.ku.dk

Prædiktion med Machine Learning

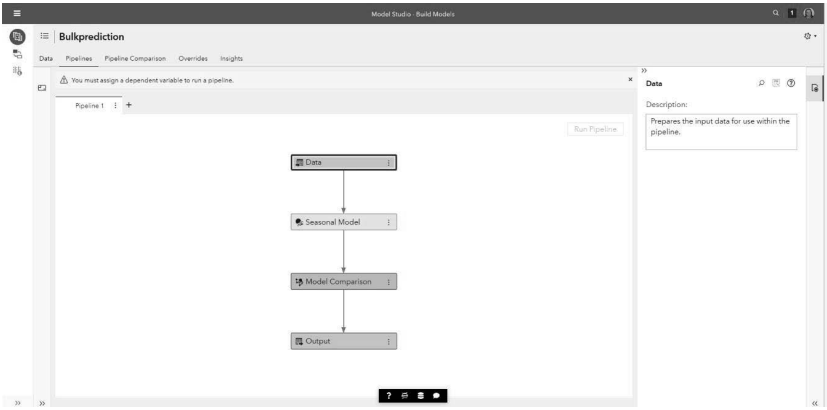
I dette eksempel forudsiges forbrugerprisindekset for to fødevarer, kalve- og oksekød samt babymad; hhv. variablene P_01_1_2_1 og P_01_1_9_3. Prædiktionerne beregnes i SAS Viya.

Babymad

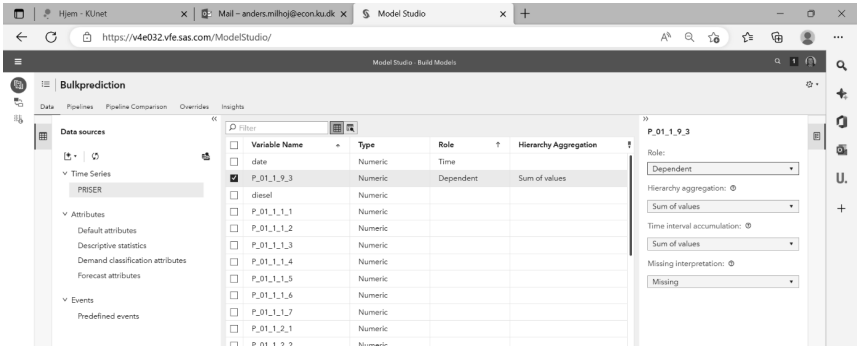
I SAS Viya skal der dannes et nyt projekt, hvilket gøres ved at udfylde en wizard. Klik



Her er valgt, at det er forecasting, projektet skal omhandle og at der skal bruges sæsonmetoder. Dernæst trykkes på Save og dernæst skal man åbne fanebladet Pipelines.



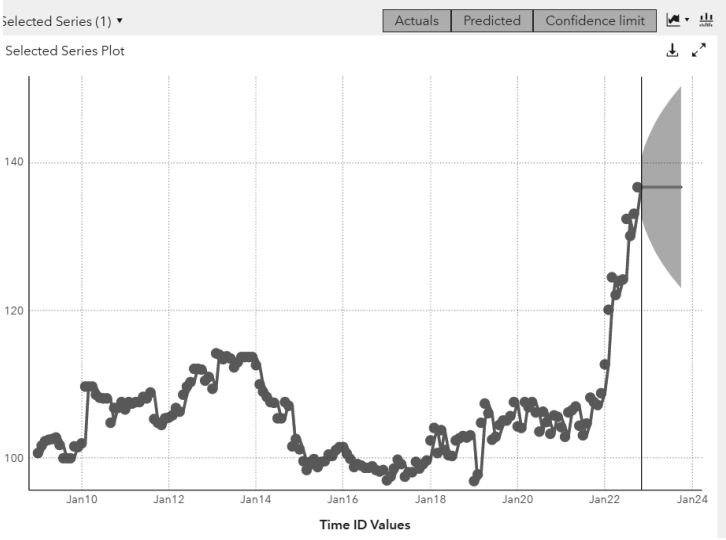
Via fanebladet data kommer man ind i datasættet og kan vælge variablen P_01_1_9_3, der er babymad, som Dependent variable.



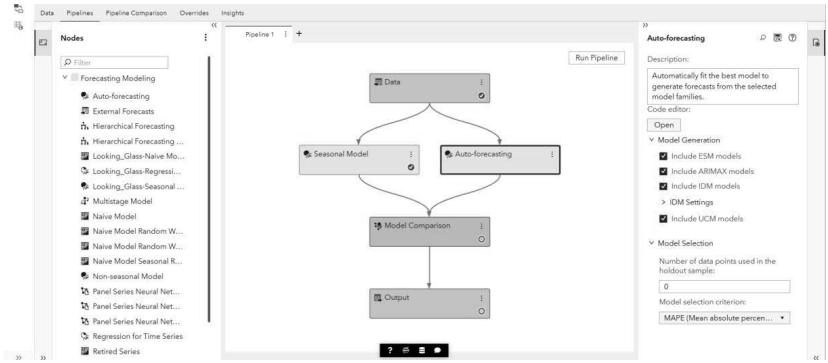
Tilbage i fanebladet Pipelines klikkes der på knappen Run Pipeline, hvorefter nogle urvisere ruller rundt mens systemet arbejder på sagen. Til sidst er der grønne tjekmarks ved alle fire kasser.

For at se resultatet, kan man fx højreklikke i kassen Seasonal Forecasting og vælge Forecast viewer. Det er tydeligt, at der er forudsagt uden en sæsonmodel, og der er valgt en simpel konstant forudsigelse, som fra en random walk model. Det er naturligvis lidt

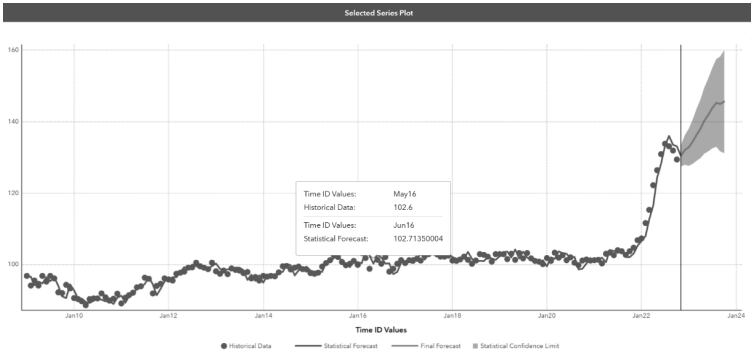
kedeligt, men mon ikke dette er, hvad data siger. Det hævdes i hvert fald, at der er afprøver "Generate forecast with seasonal ESM, ARIMAX or UCM model".



Ved at klikke på datanoden kan der tilføjes en ny node, som vælges til Auto-forecasting. I denne node kan der klikkes, så alle tænkelige forudsigelsesmetoder benyttes. I Forecast viewer kan det ses, at den valgte model er simpel eksponentiel udglatning.



For okse- og kalvekød er den tilsvarende forudsigelse lidt mere interessant, idet den opadgående trend de første 10 måneder af 2022 forventes at forsætte om end ikke monotont.



Bulkprædiktion med SAS procedurer

I dette afsnit forudsiges prisen for 48 forbrugerprisindeks for fødevarer med automatiserede kald af SAS-procedurer.

Der er flere SAS-procedurer, der kan danne forudsigelser af tidsrækker, hvoraf to uden videre kan håndtere flere tidsrækker i et kald, mens andre skal pakkes ind i en SAS-makro, der kalder proceduren for hver tidsrække for sig.

Proc ESM

Da det på forhånd er uklart, om der er sæson og trend i de enkelte prisindeksserier, anvendes (Holt-)Winters metode, der kan tage højde for både sæson og trend. Selve metoden er en simpel algoritme, der iterativt kombinerer den bestemte komponent (niveau, trend og sæson) med den nyeste observation ved hjælp af tre vægte, der estime-

$$\tilde{X}_t = \alpha \frac{X_t}{S_{t-12}(May)} + (1-\alpha)\left(\tilde{X}_{t-1} + b_{t-1}\right),$$

$$b_t = \gamma\left(\tilde{X}_t - \tilde{X}_{t-1}\right) + (1-\gamma) b_{t-1}$$

$$S_t(May) = \omega \frac{X_t}{\tilde{X}_t} + (1-\omega) S_{t-12}(May).$$

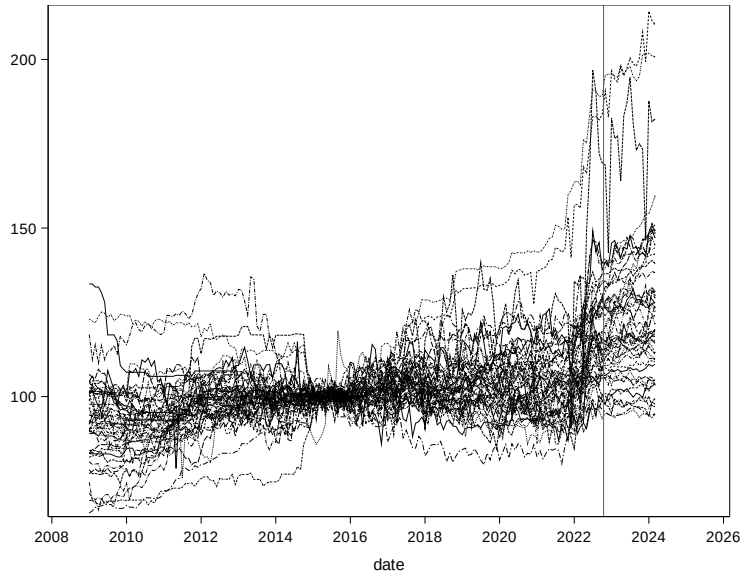
res ved at minimere kvadratafvigelsessummen af forudsigelsesfejlene i observationsperioden. Forudsigelsen beregnes ved at gentage formlen for den seneste observation fx ved successivt at addere det sidst fundne trendbidrag.

De tre vægte skal ligge mellem nul og en, i praksis begrænses de af numeriske grunde til værdier mellem 0.001 og 0.999. En værdi nær ved en lægger mest vægt på den nyeste observation, mens en værdi tæt ved nul tillægger fortiden stor vægt i beregningerne. For tidsrækker med sæson er værdien erfaringsmæssigt meget lille, dvs. lig med eller kun lidt over den nedre grænse på 0.001; mens værdier væsentligt over 0.001 kun dukker op i tidsrækker med en meget svag eller slet ingen sæson.

Proceduren kaldes med den viste kode. I `forecast` statementet angiver kolonnet i `P:`, at proceduren skal køres for alle variable, der begynder med P i variabelnavnet. Forudsigelserne gemmes i datasættet `forudsigelser`.

```
proc esm data=b.priser plot=nonell print=none lead=17
OUT=forudsigelser OUTEST=parametre;
id date interval=month;
forecast P:/method=winters;
run;
```

Datasættet `parameter` indeholder de estimerede vægte. Ingen af vægtene er under 0.1 og kun syv er under 0.2, mens syv antager den højeste værdi 0.999 og nitten er over 0.5. Der betyder alt i alt, at der ikke er synderlig sæson i de fleste af de 48 prisrækker.



Proc X12 – X13

Disse procedurer til sæsonrensning kan bestemme en sæson ARIMA model automatisk. I forbindelse med sæsonrensningen bruges denne model til at forecaste og backcaste tidsrækkens værdier, så de mange glidende gennemsnit i X11's sæsonrensningsalgoritme fungerer bedre for de første og sidste år af tidsækken. Desuden indgår ARIMA modellen centralt i SEATS metoden til sæsonrensning, idet denne metode bestemmer vægtene til de glidende gennemsnit ud fra den fittede model. Så selvom proceduren egentlig er tiltænkt andre formål, bestemmer den en sæson ARIMA model.

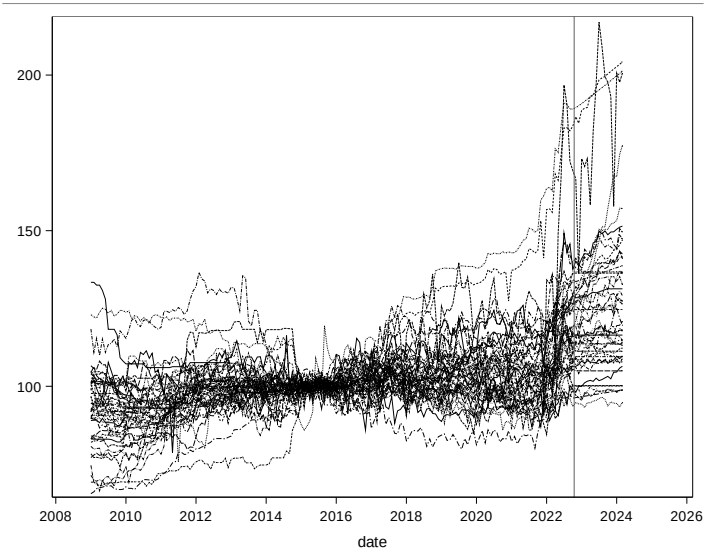
Når der forudsiges mange tidsrækker på én gang, kan man gå i detaljer i analysen af enkelt tidsrække. Derfor anvendes en opsætning, der forventes at fungere nogenlunde for alle rækker. I dette tilfælde anvendes den multiplikative sæsonrensningsmetode, idet modellen dog estimeres på de logaritmisk transformerede data. De to `ods output` statements sikrer hhv., at der dannes et datasæt, `predicted`, der indeholder forudsigelserne, og et datasæt `ARIMA_model`, der indeholder oplysninger om, hvilken sæson ARIMA model, der bestemmes. Kolonnet i `var` statementet, `P:`, siger at proceduren kører på alle variable, hvis navn begynder med P i datasættet. Det er denne facilitet, der gør det let at benytter Proc X12 til bulkforudsigelser.

```
proc X12 data=b.priser date=date plots=none;
var P:;
transform function=log;
x11;
automdl;
forecast lead=17;
ods output ForecastCL=predicted;
ods output FinalModelChoice=ARIMA_model;
run;
```

Modellen udvælges ved en afvejning af, at residualvariansen skal være lille og at testet for restautokorrelation i residualerne godkender bedst muligt. Desuden sammenlignes med airline-modellen ARIMA(0,1,1) × $ARIMA_{12}$(0,1,1), som efter det meget succesfulde eksempel i Box & Jenkins(1976) fungerer som default model for tidsrækker med sæson; denne model foretrækkes ofte, se tabellen.

I dette eksempel er det ikke så oplagt, at det overhovedet er sæson at spore i forbrugerprisindeks, for hvorfor skulle der dog være sæson i prisen på babymad? Men i de vælges bare en model uden sæson, hvad den gør for 14 af de i alt 48 fødevarer, hvor der slet ikke indgår sæsonparametre. Tabellen viser at de simple ikke sæson modeller ARIMA(1,1,0) og ARIMA(0,1,1) tilsammen vælges for 11 varer.

Model	*Antal*
ARIMA(0,1,1) x ARIMA12(0,1,1)	17
ARIMA(0,1,1) x ARIMA12(0,0,0)	6
ARIMA(1,1,0) x ARIMA12(0,0,0)	5
ARIMA(0,1,1) x ARIMA12(1,0,0)	3
ARIMA(1,1,0) x ARIMA12(1,0,0)	3
ARIMA(1,1,0) x ARIMA12(0,0,1)	2
ARIMA(1,1,0) x ARIMA12(0,1,1)	2
ARIMA(0,1,0) x ARIMA12(1,0,1)	1
ARIMA(0,1,1) x ARIMA12(0,0,1)	1
ARIMA(0,2,1) x ARIMA12(0,0,0)	1
ARIMA(1,1,0) x ARIMA12(1,0,1)	1
ARIMA(1,1,1) x ARIMA12(0,0,0)	1
ARIMA(1,1,2) x ARIMA12(0,1,1)	1
ARIMA(1,2,1) x ARIMA12(0,1,1)	1
ARIMA(2,1,0) x ARIMA12(0,0,0)	1
ARIMA(2,1,0) x ARIMA12(0,1,1)	1
ARIMA(3,0,0) x ARIMA12(1,0,1)	1



Proc Varmax

Denne procedure er tiltænkt estimation af flerdimensionale tidsrækker, men den er også i stand til at bestemme en ARMA(p,q) model automatisk, altså bestemme p og q, dvs. hhv. antal autoregressive og glidende gennemsnits parametre. Desværre bestemmes antal differensdannelser ikke af proceduren, og den kan heller ikke håndtere sæsonmodeller på en meningsfuld måde.

I denne procedure kan notationen med et kolon ikke anvendes, så derfor pakkes den viste kode ind i en SAS-makro, der successivt kalder Proc Varmax for hver af de 48 tidsrækker. Outputdatasættene samles til sidst i datasæt med hhv. modelvalget og forudsigelserne.

I den anvendte modelklasse anvendes en differensdannelse. Det skyldes, at en stationær model uden en differens ville tvinge forudsigelserne til at konvergere mod gennemsnittet af prisindekset i hele observationsperioden, hvilket ville være meningsløst i sammenhængen. En yderligere effekt af differensen er, at den estimerede middelværdi af tidsrækken efter differensen danner et trendbidrag, der vil kunne bygge et sandsynligvis stigende prisniveau ind i forudsigelserne, hvad der kan være rimeligt for mange af rækkerne.

I optionerne angives, at alle værdier af antal autoregressive led kan være p=0, 1, .., 14. Et højt p kan derved tage højde for eventuelle sæsoneffekter i de månedlige indeks.

Sæson kan dog også opnås for lave ordner, p, hvis der er komplekse rødder i det autoregressive polynomium, men det er umuligt for p = 0 og p =1.

```
PROC VARMAX data=b.priser;
id date interval=month;
model P_01_1_9_3/method=ml dif=(P_01_1_9_3(1))
minic=(type=aic p=14 q=0);
ods output modeltype=model;
output out=pred_ lead=17;
run;
```

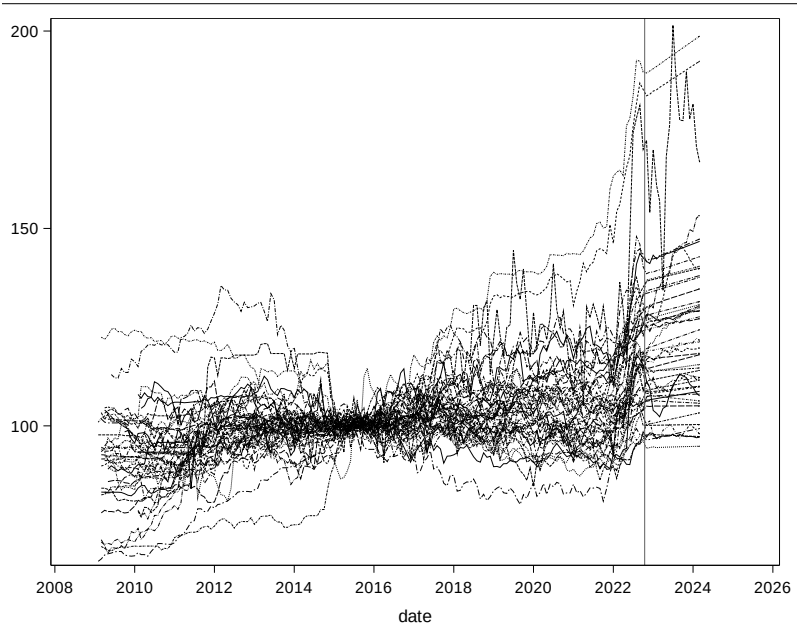
I den viste kode anvendes Akaikes informationskriterium, som udvælger følgende modeller

Model	*Antal*
AR(1)	16
AR(2)	9
AR(3)	6
AR(12)	5
AR(0)	3
AR(13)	2
AR(14)	2
AR(8)	2
AR(11)	1
AR(4)	1
AR(6)	1

Anvendes Schwarts' bayesianske informationskriterium, BIC, straffes der mere for antal parametre, hvad der tydeligt ses af tabellen, hvori det ses, at der kun er to tidsrækker med flere end tre autoregressive parametre. Hele 33 af de 48 tidsrækker har højst orden et, dvs. at der ikke kan være komplekse rødder i det autoregressive polynomium, så modellen kan på ingen måde beskrive sæsoneffekter.

Model	*Antral*
AR(1)	20
AR(0)	13

Model	*Antral*
AR(2)	10
AR(3)	3
AR(4)	1
AR(6)	1



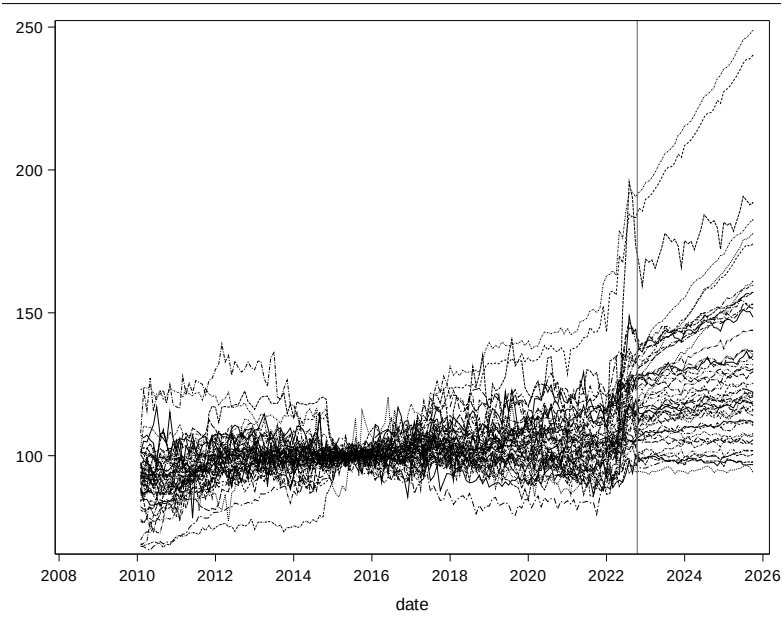
Proc UCM

Denne procedure estimerer modeller for uobserverede komponenter for en tidsrække. I det viste program anvendes et niveau og en hældningskoefficient, der tillades at variere med en varians, der estimeres i proceduren. Sæsonkomponenten holdes konstant i hele tidsperioden, idet dens varians sættes til nul. Signifikansen af de tre tilpassede komponenter for den sidste observation angives i en af outputtabellerne. De nyeste tilpassede komponenter benyttes ved beregning af forudsigelserne.

I denne procedure kan notationen med et kolon ikke anvendes, så derfor pakkes den viste kode ind i en SAS-makro, der successivt kalder Proc UCM for hver af de 48 tidsrækker. Outputdatasættene samles til sidst i et enkelt datasæt for hhv. komponentsignifikansen og forudsigelserne.

```
PROC UCM data=b.priser;
id date interval=month;
model &var;
level;
slope;
season length=12 var=0 noest;
outlier;
estimate;
forecast outfor=pred_&var lead=36;
ods output ComponentSignificance=significance_&var;
run;
```

I 34 af de 48 tidsrækker er sæsonkomponenten signifikant på et 5% niveau, mens trendkomponenten kun er signifikant i 11 af de 48 tidsrækker.



Interpolation af vægt for spædbørn

Sören Möller[1,2] og Gitte Zachariassen[1,3]

[1] Klinisk Institut, Syddansk Universitet
[2] Open Patient data Explorative Network, Odense Universitetshospital
[3] H.C. Andersen Børne- og Ungehospital, Odense Universitetshospital

Introduktion

I mange kliniske studier på nyfødte er der behov for at sammenligne deres vægt og højde med den vægt og højde, der forventes ud fra referencedata på befolkningsniveau. Da små børn vokser hurtigt og ikke-lineært, og børnene typisk måles på varierende dage, der ikke stemmer overens med de tidspunkter, der rapporteres i referencedata, er der behov for at interpolere referencedata til de tidspunkter, hvert enkelt barn er blevet målt på, specielt med henblik på at bestemme z-scores på vægt og højde, som er udbredte udfaldsmål i denne type studier. Endvidere kan der være behov for en sekundær interpolation, for at kunne sammenligne børn, der er målt på forskellige tidspunkter, indenfor samme studie. Vi vil i dette bidrag beskrive den parametriske interpolationsmodel, vi med succes har brugt på flere kliniske studier i Odense, samt diskutere statistiske overvejelser relateret til disse interpolationer.

Problemstilling og formål

Vi ønsker en model, der baseret på referencedata, svarende til gennemsnit og spredning af vægt på bestemte aldre kan interpolere til mellemliggende aldre med det formål at bestemme z-scores for vægten på de tidspunkter et barn er blevet målt.

For at den ønskede metode kan anvendes i praksis af kliniske forskere er der nogle kriterier til metoden vi ønsker at have opfyldt:

- Metoden skal kunne anvendes på nye referencedatasæt uden adgang til rådata for referencedatasættet, så det, der typisk publiceres, skal være nok information.
- Metoden skal kunne anvendes på forskellige aldersintervaller, ved at bruge forskellige referenedata.
- Metoden ønskes at være parametrisk, sådan at den kan afrapporteres som eksplicitte formler der kan anvendes på nye studier.
- Metoden skal være kompatibel med efterfølgende interpolation på z-værdiskalaen for at bestemme børns værdi mellem målepunkter.

Vi præsenterer i dette bidrag vores forslag til en algoritme, der polynomielt interpolerer referencedata, for at betsemme z-scores på de tidspunkter, hvor et barns vægt er blevet målt i et sundhedsvidenskabeligt studie. Denne algortime har vi tidligere brugt i et antal forskningsprojekter [2,3,4,1].

Foreslået algoritme

Vi tilpasser polynomier til henholdsvis gennemsnitsvægten og en spredning under gennemsnitsvægten på referencedata for at opnå en parametrisk model for vægtudviklingen over tid:

$$W_m(T) = \sum_{k=0}^{K} b_{k,m} T^k$$

$$W_{m-s}(T) = \sum_{k=0}^{K} b_{k,m-s} T^k$$

hvor T er barnets alder (henholdsvis gestationsalder for nyfødte og alder i måneder eller år for større børn). Ud fra dette bestemmer vi så den forventede gennemsnitsvægt og spredning ved en given, vilkårlig, alder t og omregner en observeret vægt w til en z-score $z(t, w)$ for det enkelte observerede barn:

$$z(t, w) = \frac{w - W_m(T)}{W_m(t) - W_{m-s}(t)}.$$

Evaluering af algoritmen

Vi evaluerer vores algoritme på to forskellige referencedatasæt:

- Svenske referencedata på nyfødtes vægt, dækkende over gestationsalder 1¨68 til 555 dage [5].
- WHOs referencedata for små børns vægt, dækkende over alder 0 til 60 måneder [6].

Begge referencedata er stratificeret på køn, derfor er vores evaluering også stratificeret på køn. For hver af de fire datasæt har vi tilpasset den polynomielle model med grad 1 til 6, bestemt overensstemmelse mellem referencedata og modelprædiktionen på de tidspunkter, der er inkluderet i referencedata, og grafisk fremstillet både referencekurve og afvigelse mellem model og reference.

Resultater

Vi observerer på figuren af de nyfødtes vægt fra de svenske referencedata (Figur 1) at første til tredje grads polynomier giver tydelige afvigelser fra de observerede data, men at fjerde til sjette grads polynomier giver en overenstemmelse med referencedata med kun meget begrænsede afvigelser, som vi vurderer at være små nok, til ikke at være klinisk relevant. Det er desuden tydeligt, at modellerne divergerer markant fra hinanden når vi kommer ud over tidspunktet for den sidste referencemåling ved 555 dages gestationsalder. Dette er forventet, men understreger at modellen ikke bør bruges til at ekstrapolere børns vægt uden for referencedatas tidsinterval.

For WHOs referencedata (Figur 2) for børn op til fem års alderen får vi lignende resultater dog med mindre afvigelser også for de lavere ordens polynomier. Fra 6 måneders alder og opefter er anden til sjette grads polynomier acceptable, men stadig med en tydelig (dog måske ikke klinisk relevant) fordel for polynomier af femte og sjette grad.

Generelt var der kun yderst begrænsede kønsforskelle i modellernes opførsel og resultaterne for en spredning under gennemsnit var generelt meget lig resultaterne for gennemsnittet.

Diskussion

Overordnet kan vi konkludere at vores polynomielle model giver god overenstemmelse med referencedata, såfremt man vælger et polynomium af grad fem eller seks. I den praktiske anvendelse er det dog vigtigt at huske, at modellens koefficienter er specifikt for hver reference, og derfor skal tilpasses på ny, hvis referencen skiftes. Endvidere bør modellen kun anvendes for at bestemme z-værdier i det tidsinterval referencen dækker, og ikke ektrapolere uden for dette.

Referencer

1. Ahnfeldt, A. M., Hyldig, N., Li, Y., Kappel, S. S., Aunsholdt, L., Sangild, P. T., and Zachariassen, G. FortiColos - a multicentre study using bovine colostrum as a fortifier to human milk in very preterm infants: study protocol for a randomised controlled pilot trial. *Trials 20*, 1 (May 2019), 279.
2. Kappel, S. S., Sangild, P. T., Ahnfeldt, A. M., ttir, V., Soernsen, L. J., Bak, L. B., Friborg, C., ller, S., Zachariassen, G., and Aunsholt, L. A Randomized, Controlled Study to Investigate How Bovine Colostrum Fortification of Human Milk Affects Bowel Habits in Preterm Infants (FortiColos Study). *Nutrients 14*, 22 (Nov 2022).
3. Klamer, A., Toftlund, L. H., Grimsson, K., Halken, S., and Zachariassen, G. IQ Was Not Improved by Post-Discharge Fortification of Breastmilk in Very Preterm Infants. *Nutrients 14*, 13 (Jun 2022).
4. lholm Hansen, B., Cueto, H., Padkaer Petersen, J., Zachariassen, G., nderby Christensen, P., Breindahl, M., ler Kesmodel, U., and Brink Henriksen, T. years. *Acta Paediatr 111*, 9 (Sep 2022), 1695–1700.
5. Niklasson, A., and Albertsson-Wikland, K. Continuous growth reference from 24th week of gestation to 24 months by gender. *BMC Pediatr 8* (Feb 2008), 8.
6. World Health Organization. WHO child growth standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: methods and development. *WHO* (2006).

Fig. 1: Evaluering af algoritmen på de svenske referencedata

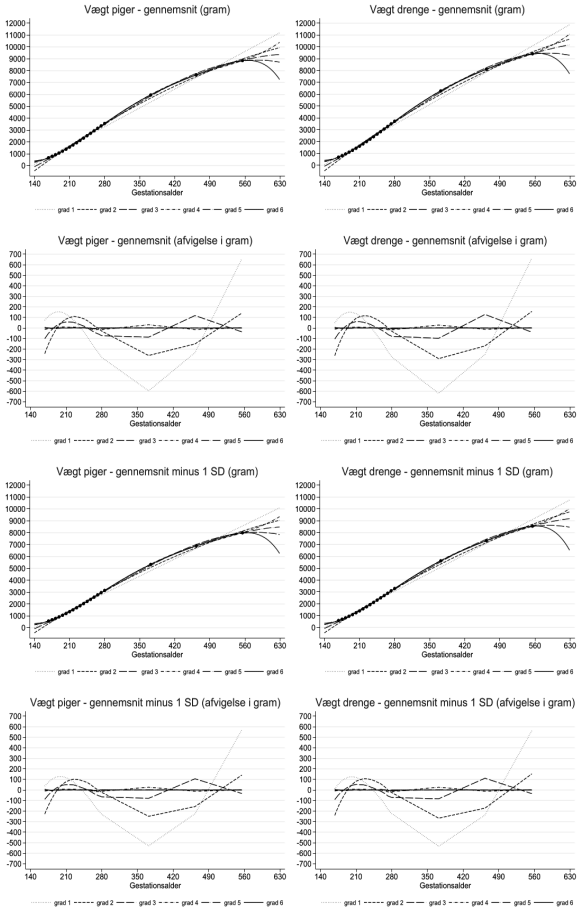
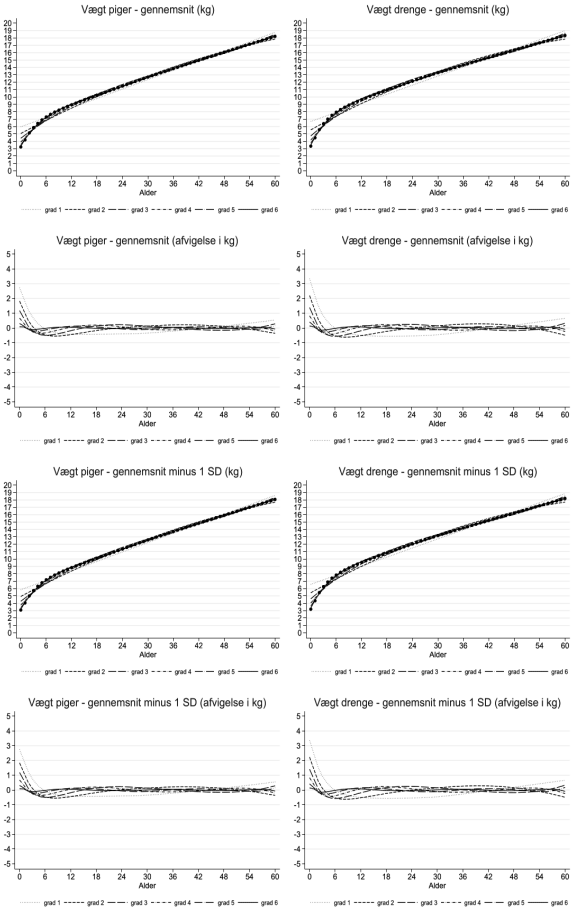


Fig. 2: Evaluering af algoritmen på WHOs referencedata



Examining sibship constellation and risk of multiple sclerosis

– an example of register-based research at its best

By Klaus Rostgaard

Danish Cancer Society Research Center, Danish Cancer Society, Copenhagen

klar@cancer.dk

Abstract

Epstein-Barr virus infection, and perhaps almost exclusively delayed Epstein-Barr virus infection, seems a prerequisite for multiple sclerosis development. Siblings provide protection against infectious mononucleosis by occasionally preventing delayed primary Epstein-Barr virus infection with its associated high risk of infectious mononucleosis. Each additional sibling provides further protection according to the age difference between the index child and the sibling. The closer in age, the higher protection and with younger siblings being more protective against infectious mononucleosis than older siblings. If the hypothesis that delayed Epstein-Barr virus infection is necessary for the development of multiple sclerosis is true, then the relative risk of multiple sclerosis as a function of sibship constellation should mirror the relative risk of infectious mononucleosis as a function of sibship constellation. Such an indirect hypothesis test is necessitated by the fact that age at primary Epstein-Barr virus infection will be unknown for practically all people not having experienced infectious mononucleosis. According to a recent paper (Rostgaard et al. 2022) this null hypothesis seems to be fundamentally true – with the hopeful perspective that future multiple sclerosis can essentially be eradicated by early Epstein-Barr virus vaccination. In this talk I shall discuss the statistical modelling in that paper and why it took so long to reach this conclusion.

Introduktion

Det er for nyligt blevet sandsynliggjort at en nødvendig forudsætning for at udvikle sclerose er at man groft sagt inficeres første gang med Epstein-Barr virus (EBV) som teenager eller senere.[1] Implikationen er, at det derfor burde være muligt at forhindre fremtidige tilfælde af sclerose ved at vaccinere med EBV i en tidlig alder. Den konklusion kunne man være kommet snublende nær for længe siden, men ingen gjorde det. Her vil vi prøve at forklare hvorfor det ikke skete tidligere. Selve modelarbejdet, resultaterne og den formodede biologi bag er behandlet i Rostgaard et al.[1] og referencer deri.

Langt de fleste smittes med EBV på et tidspunkt, og når man bliver smittet er man smittet for evigt. De fleste bliver smittet i en af to bølger: enten som 0-2 årige eller som teenagere. EBV overføres gennem spyt, og er generelt ikke særligt smitsomt. Når man smittes første gang med EBV kan man få kyssesyge (infektiøs mononukleose). Det er

en overreaktion fra immunsystemet på infektionen.[2] Typiske symptomer er træthed, feber, hovedpine, hævede lymfeknuder og halsbetændelse.[3] Kyssesyge er en hyppig følgevirkning af primær EBV-infektion i teenage-årene, men ret sjælden ved tidligere infektion.[2] Derfor er kyssesyge reelt en markør for sen (teenage-årene eller senere) primær EBV infektion. Man har længe vidst at ca. halvdelen af befolkningen i vestlige lande er EBV-naive ved indgangen til teenage-årene, men det er særligt pædagogisk og elegant demonstreret i et modelarbejde fra 2019.[2]

Den forsinkede hypotese

Man har længe forbundet EBV med sclerose, og ideen om sen EBV-infektion som risikofaktor for sclerose går tilbage ihvertfald til midten af 1990'erne.[4,5] Men de data de fleste har haft til rådighed angående alder ved primær EBV-infektion har netop været data om personer der har haft kyssesyge, eller som er blevet testet for det. Så de pågældende forfattere har været tilbageholdende med at skelne meget skarpt mellem kyssesyge og forsinket EBV-infektion og med at fordele betydning af sen EBV-infektion i sig selv, og kyssesyge som en ekstra komplikation ved primær EBV-infektion. Nogen burde dog have fået ideen om tingenes rette sammenhæng efter at en række store studier[6] (Haarh 1995 OR=2.80, Hernan 2001 OR=2.20, Nielsen 2007 OR=2.27, Zaadstra 2008 OR=2.22, Ramagopalan 2009 OR=2.06, Ahlgren 2009 OR=2.06) ret konsistent fandt en relative risikoforøgelse for sclerose efter kyssesyge på ca 2. Dette burde være blevet fortolket som en mulig konsekvens af, at den reelle (men uobserverbare) andel under risiko for at få kyssesyge var omtrent halveret når personerne startede med at være teenager.[5,7–9] Det burde også have stået klart, både ud fra hverdagserfaring og fra literaturen,[2,10] at hvis kyssesyge var en nødvendig betingelse for at få sclerose skulle den relative risiko for sclerose efter kyssesyge være meget højere, svarende til at man havde selekteret sig ind i en meget mindre population der reelt var under risiko.

I epidemiologien har man ofte anvendt søskendeflokkarakteristika (fx birth order, antal søskende, antal ældre og yngre søskende, aldersafstand til nærmeste søskende o.s.v.) som proxier for infektiøs eksponering. For at belyse om vores ideer om smittespredning i familier og infektionsmønstre holdt vand for EBV og kyssesyge gennemførte vi i 2014 et simpelt, men meget detaljeret studie af søskendeflokkarakteristika som prediktor for kyssesyge.[11] Studiet estimerede både direkte effekter (en tydeligt forøget risiko for at få kyssesyge når man havde nogle 0-3 årige søskende) og en indirekte effekt (at det var beskyttende at have især yngre søskende, at hver søskende betød mere beskyttelse, og at den beskyttende effekt var størst når aldersdifferencen var mindst). Det gav anledning til en fortolkelig model hvor man især smittes med EBV af søskende i alderen 0-3 år, og hvis man ikke reagerer med kyssesyge på det tidspunkt vil man således være vaccineret mod at få kyssesyge senere i livet. Studiet var baseret på hospitals-kontakter med kyssesyge som udfald, men vi kunne I samme studie ret nøjagtigt genfinde vores primære (indirekte) effektestimater blandt bloddonorer med selvrapporteret kyssesyge som udfald.[11] Så vi havde dermed en model der hang nogenlunde logisk og troværdigt sammen

og som kunne bruges til at prediktere fravær af risiko for sen EBV-infektion (og dermed kyssesyge). Dermed var den sidste logiske forudsætning på plads for at kunne formulere hypotesen om sen EBV-infektion som nødvendig i udviklingen af sclerose, som en statistisk hypotese om at den indirekte (beskyttende) effekt af søskendeflok-sammensætning på hazard ration var den samme for de to udfald kyssesyge og sclerose. Der fandtes på det tidspunkt studier der allerede 10 år tidligere havde vist den beskyttende effekt af søskende, og især yngre søskende på risikoen for sclerose, men de var ikke så detaljerede og statistisk præcise.[12–14] Dette er omvendt blandt andet en følge af at vi kun kender slægtningene i de yngre kohorter, og at det tager tid (alder) at få sclerose. Vi har fx 4 gange så mange sclerose-tilfælde i dag som Bager *et al.* havde 15 år tidligere.[1,14]

Og hvad skal der så ske?

Det eneste der ikke umiddelbart passer med vores model er, at man også kan få kyssesyge før teenage-alderen, og i så tilfælde har man den samme forhøjede relative risiko for sclerose (SIR≈2), som hvis man havde fået kyssesyge senere. Vi har som løsning på det problem postuleret at de biologiske processer som starter udviklingen frem mod sclerose er de samme blandt pre-teenagere og teenagere, men bare meget mere hyppige blandt teenagere.[2] Så den eneste umiddelbare svaghed ved vores model kan forhåbentlig blive det vink der gør det muligt entydigt at identificere hvad det er for biologiske processer, der er tale om. Der er mange hypoteser på området, men også stor usikkerhed.[1] At identificere disse biologiske processer må antages at være en nødvendighed for at kunne skabe en nødvendig videnskabelig konsensus,[15] som grundlag for at udvikle en passende vaccine og vaccinestrategi, selv om ikke alle mener det er nødvendigt.[16] Bl.a. fordi en fornuftig vaccinestrategi blot vil sørge for "kunstigt" at genskabe det forhold mellem mennesker og virus som har eksisteret gennem årtusinder, hvor langt de fleste inficeres med EBV i de første leveår, uden at det har nogen væsentlige følgevirkninger.[1,17–19]

References

1. Rostgaard, K., Nielsen, N. M., Melbye, M., Frisch, M. & Hjalgrim, H. Siblings reduce multiple sclerosis risk by preventing delayed primary Epstein-Barr virus infection. *Brain* (2022). doi:10.1093/brain/awac401
2. Rostgaard, K. *et al.* Primary Epstein-Barr virus infection with and without infectious mononucleosis. *PLoS One* **14**, e0226436 (2019).
3. Balfour, H. H., Dunmire, S. K. & Hogquist, K. A. Infectious mononucleosis. *Clin. Transl. Immunol.* **4**, e33 (2015).
4. Haahr, S., Koch-Henriksen, N., Møller-Larsen, A., Eriksen, L. S. & Andersen, H. M. Increased risk of multiple sclerosis after late Epstein-Barr virus infection: a historical prospective study. *Mult. Scler.* **1**, 73–7 (1995).
5. Haahr, S., Plesner, a. M., Vestergaard, B. F. & Höllsberg, P. A role of late

Epstein-Barr virus infection in multiple sclerosis. *Acta Neurol. Scand.* **109**, 270–275 (2004).

6. Handel, A. E. *et al.* An updated meta-analysis of risk of multiple sclerosis following infectious mononucleosis. *PLoS One* **5**, (2010).
7. Lai, P. K., Mackay-Scollay, E. M. & Alpers, M. P. Epidemiological studies of Epstein-Barr herpesvirus infection in Western Australia. *J. Hyg. (Lond).* **74**, 329–337 (1975).
8. Hesse, J., Ibsen, K. K., Krabbe, S. & Uldall, P. Prevalence of antibodies to Epstein-Barr virus (EBV) in childhood and adolescence in Denmark. *Scand. J. Infect. Dis.* **15**, 335–338 (1983).
9. Kangro, H. O. *et al.* Seroprevalence of antibodies to human herpesviruses in England and Hong Kong. *J. Med. Virol.* **43**, 91–96 (1994).
10. Hjalgrim, H. On the aetiology of Hodgkin lymphoma. *Dan. Med. J.* **59**, B4485 (2012).
11. Rostgaard, K. *et al.* Sibship structure and risk of infectious mononucleosis: a population-based cohort study. *Int. J. Epidemiol.* **43**, 1607–1614 (2014).
12. Montgomery, S. M., Lambe, M., Olsson, T. & Ekbom, A. Parental age, family size, and risk of multiple sclerosis. *Epidemiology* **15**, 717–23 (2004).
13. Ponsonby, A.-L. *et al.* Exposure to infant siblings during early life and risk of multiple sclerosis. *JAMA* **293**, 463–9 (2005).
14. Bager, P. *et al.* Sibship characteristics and risk of multiple sclerosis: A nationwide cohort study in Denmark. *Am. J. Epidemiol.* **163**, 1112–1117 (2006).
15. Oreskes, N. *Why trust science?* (Princeton University Press, 2019).
16. Jons, D., Sundström, P. & Andersen, O. Targeting Epstein-Barr virus infection as an intervention against multiple sclerosis. *Acta Neurol. Scand.* **131**, 69–79 (2015).
17. Morgan, A. & Khanna, R. Epstein-barr virus vaccines. in *Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis* (eds. Arvin, A. et al.) (Cambridge University Press, 2007).
18. Cruz-Muñoz, M. E. & Fuentes-Pananá, E. M. Beta and Gamma Human Herpesviruses: Agonistic and Antagonistic Interactions with the Host Immune System. *Front. Microbiol.* **8**, 2521 (2017).
19. Niederman, C. & Evans, A. Epstein-Barr Virus. in *Viral Infections of Humans. epidemiology and Control.* (eds. Evans, A. & Kaslow, R.) 253–283 (Plenum Medical Book Company, 1997).

Using home-scan data to analyze dietary changes in relation to major life changing events

Sinne Smed, Department of food and resource economics, University of Copenhagen

Based on a range of papers, combining detailed home-scan food purchase data from an unbalanced panel of 12000 households (20000 individuals) with official high-quality individual register-data, we discuss the possibilities of using this data combination to study the effects of life-changing event on dietary health.

This structure allows us to, 1) consider dynamic adjustments to life-changing events. Unemployment and retirement lead e.g. to some short-run changes, probably caused by substitution from out-of-home to in-house consumption, whereas real dietary changes, are found in the longer run. Life-style related illnesses on the other hand lead to some short-run changes that, in most cases, returns to before diagnose levels in the long-run, 2) consider socio-demographic and attitude based differences in the response to life changing events. There is e.g. differences in the dietary adjustment to retirement dependent on whether the individual retire from work or from a position outside the labor market and due to family status.

One disadvantage of using this dataset to analyze dietary behaviour is that we observe food purchases of households, not individual food consumption. Another disadvantage is that the data are observational and not directly constructed with the aim of analyzing dietary behaviour, hence there is considerable amount of noise in the data and we are in some cases challenged by small samples and also endogeneity.

Hvor stor en andel af 'signifikante' resultater er falske?

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

December 2022

Abstract: *En stor del af publicerede statistisk signifikante effekter og sammenhænge er falske, dvs. et resultat af fejlagtige H_o afvisninger. Udover almindelig data-mining, skyldes det (mindst) tre forhold. For det første, at sandsynligheden for H_o, givet at H_o afvises, er højere end Type-I fejlsandsynligheden. For det andet, at sandsynligheden for H_o, givet data, er højere end p-værdien. Og for det tredje, at der sjældent korrigeres for multiple tests. Der er behov for en kulturændring i vores empiriske praksis.*

Keywords: *p-værdi; 5% signifikansniveau; false discovery rate; Bayes formel; sandsynligheden for H_o; multiple tests.*

1. Indledning.

Det er velkendt, at der er en replikationskrise i empiriske videnskaber. Mange publicerede resultater har vist sig ikke-replikerbare i nye stikprøver, og det gælder både for resultater baseret på kontrollerede eksperimenter og for studier baseret på ikke-eksperimentelle, passivt observerede data. Årsagerne er mange: data-mining, p-hacking, uhensigtsmæssige incitamentsstrukturer i det akademiske system, etc. Men en medvirkende årsag er endvidere, at de

[1]Tak til Jesper Schneider for meget udbytterige samtaler om statistisk metodologi. Engsted and Schneider (2022) indeholder en detaljeret diskussion af de særlige problemer som passivt observerede ikke-eksperimentelle data indenfor samfundsvidenskaberne giver i relation til det klassiske hypotesetest-setup.

konventionelle statistiske metoder vi anvender, har en indbygget tendens til at finde for mange 'signifikante' resultater.

For det første er *false discovery* raten, dvs. sandsynligheden for at H_o er sand hvis H_o forkastes, $P(H_o \text{ er sand} \mid H_o \text{ forkastes})$, ofte meget højere end signifikansniveauet (Type-1 fejlsandsynligheden), dvs. den omvendt betingede sandsynlighed $P(H_o \text{ forkastes} \mid H_o \text{ er sand})$, især hvis styrken af testet er lav. For det andet er det konventionelle 5% signifikansniveau en meget lav overlægger i den forstand, at sandsynligheden for H_o, givet de observerede data, $P(H_o \mid D)$, ofte er meget højere end p-værdien $P(D+ \mid H_o)$, hvor $D+$ angiver de observerede data eller mere ekstreme data. En p-værdi på 5% vil typisk indebære, at $P(H_o \mid D)$ er væsentligt højere end 5%. Og for det tredje korrigeres signifikansniveauet og p-værdien i empiriske studier sjældent for multiple tests, hvilket indebærer, at når der udføres mange tests, vil det samlede signifikansniveau være væsentligt højere end det valgte signifikansniveau i hvert enkelt test.

I denne artikel diskuterer jeg disse forhold, og jeg argumenterer for, at det er påtrængende med en kulturændring i vores empiriske praksis.

2. Er signifikans på 5% niveau en tilstrækkelig høj overlægger?

Det traditionelle valg af et 5% signifikansniveau stammer fra den engelske statistiker Ronald Fisher, der for snart 100 år siden foreslog dette niveau ved tests i små stikprøver med eksperimentelle data (Fisher, 1925). En p-værdi mindre end 0.05 er ifølge Fisher signifikant evidens imod H_o. Det underliggende rationale for at forkaste H_o baseret på en lav p-værdi udtrykte han som følger: "*Either an exceptionally rare chance has occurred, or the theory* [modellen i H_o] *is not true*" (Fisher, 1959, p.39).

Selvom vi ofte uformelt taler om, at man ikke bør opfatte 5% som en fast regel, der altid bør benyttes uanset stikprøvestørrelse og karakteren af data iøvrigt, fungerer signifikans på 5% niveau alligevel som en de facto nødvendig betingelse for at kunne publicere et resultat. Ser man på fordelingen af teststatistikker eller p-værdier på tværs af publicerede studier, er fordelingen stort set trunkeret ved 5%. Der er kun meget få resultater med p-værdier større end 0.05, og det meste af massen i fordelingen ligger lige til venstre for 0.05 (Harvey, 2017; Andrews and Kasy, 2019). Uanset om vi vil det eller ej, så jagter vi allesammen tilsyneladende stadig signifikans på 5% niveau!

2.1 False discovery rate.
At signifikans på 5% ofte vil være en ret lav overlægger, kan illustreres ved beregning af den såkaldte '*false discovery rate*', dvs. andelen af signifikante resultater, der er falske. Antag at parameteren vi er interesseret i er μ, og at vi tester H_o: $\mu = 0$ overfor H_1: $\mu \neq 0$, sådan at forkastelse af H_o indebærer, at vi har fundet en signifikant 'effekt'. Lad α og β være henholdsvis Type-I fejlsandsynligheden (dvs. signifikansniveauet) og Type-II fejlsandsynligheden for dette traditionelle hypotesetest. I bayesiansk iklædning kan vi nu lade $P(H_o)$ være á priori sandsynligheden for at H_o er sand, og vi kan med anvendelse af Bayes formel beregne den betingede sandsynlighed $P(H_o$ er sand $\mid H_o$ forkastes), jf. eksempelvis Storey (2003):

$$P(H_o \text{ er sand} \mid H_o \text{ forkastes}) = \frac{\alpha}{\alpha + \frac{(1-\beta)(1-P(H_o))}{P(H_o)}}. \tag{1}$$

Det ses af formel (1), at *false discovery* raten afhænger af det valgte signifikansniveau (α), testets styrke ($1-\beta$) og á priori sandsynligheden $P(H_o)$. Raten er naturligt stigende i α, men bemærk også, hvordan sandsynligheden for en falsk forkastelse stiger, når testets styrke falder. Tabel 1 viser *false discovery* raten for $\alpha = 0.05$ og for forskellige værdier af β og $P(H_o)$.

	Styrke, $1-\beta$			
	0.18	0.50	0.75	1.00
$P(H_o) = 0.50$	0.217	0.091	0.062	0.048
$= 0.75$	0.455	0.231	0.167	0.130
$= 0.95$	0.841	0.655	0.559	0.487

Tabel 1: *False discovery rate* beregnet for $\alpha = 0.05$ vha. formel (1)

I empiriske studier indenfor samfundsvidenskab, er det sjældent at se diskussioner af de anvendte tests styrkeegenskaber, måske fordi vi godt ved, at de tests vi anvender ofte har ret dårlige styrkeegenskaber. I et stort metastudie finder Ioannidis et al. (2017), at det typiske empiriske studie indenfor økonomi har en styrke på blot 18%. Det ses af Tabel 1, at med en så lav styrke, er sandsynligheden for en falsk forkastelse ret høj, og under alle omstændigheder markant højere end signifikansniveauet. Så selvom vi i det klassiske test kan kontrollere Type-I fejlsandsynligheden (5%), dvs. hvor

ofte vi fejlagtigt forkaster en sand hypotese, har vi ingen kontrol over den omvendt betingede sandsynlighed $P(H_o$ er sand $\mid H_o$ forkastes).

Hvad vil være et rimeligt valg af $P(H_o)$, og dermed af $P(H_1) = 1-P(H_o)$? I sidste ende er det en subjektiv vurdering hos hver enkelt analytiker, ligesom valg af signifikansniveau i teorien også er en subjektiv vurdering (selvom vi som regel blot anvender det konventionelle 5% niveau). Berger and Sellke (1987) argumenterer for, at $P(H_o) = P(H_1) = 0.50$ er det videnskabeligt 'objektive' valg og at det sjældent vil være rimeligt at sætte $P(H_o) < 0.50$, hvorimod $P(H_o) > 0.50$ er mere rimeligt. Alternativhypotesen H_1 indeholder typisk den model vi undersøger, hvor en bestemt effekt eller sammenhæng er til stede, og vi finder støtte til modellen, hvis vi forkaster H_o. Men ofte er effekten eller sammenhængen ikke et resultat af teoretiske overvejelser før man kigger på data, men snarere et resultat af data-mining med efterfølgende teoretisk rationalisering (Gigerenzer, 2004).

Indenfor empirisk finansiering, eksempelvis, er dette et udtalt fænomen. De mange risikofaktorer og afkastprediktorer for aktiemarkedet, som er publiceret i den empiriske finansieringslitteratur, er i høj grad et resultat af data-mining. Som Harvey (2017, p.1417) skriver: "*Among the many variables that researchers have explored, how many do we believe have 1:1 odds of being true return predictors before we look at the data? Very few.*" Harvey argumenterer følgelig for, at $P(H_o)$ bør sættes noget højere end 0.50. Tabel 1 viser, at med $P(H_o) > 0.50$, kan *false discovery* raten være ganske høj. Harvey (2017, p.1399) konkluderer, at den empiriske forskning producerer "*an embarrassing number of false positives - effects that will not be repeated in the future.*" Det er nærliggende at tro, at noget tilsvarende gælder for de øvrige samfundsvidenskaber.[2]

Indenfor visse fagområder er datamængden ganske betragtelig, hvorved manglende styrke i de anvendte tests ikke er et problem. Eksempelvis analyser på mikro- og registerdata, eller højfrekvente finansielle data, hvor der ofte indgår tusindvis eller millionvis af observationer. Det ses i Tabel 1, at med maksimal styrke (100%), er *false discovery* raten tæt på signifikansniveauet (5%) med neutrale á priori odds for H_o. For $P(H_o) > 50\%$ er raten dog naturligvis stadig større end 5%. I meget store stikprøver opstår et andet problem, nemlig at enhver lillebitte - og i realiteten helt ubetydelig - afvigelse fra H_o bliver signifikant på de konventionelle signifikansniveauer.

[2]Ifølge Benjamin et al. (2018) vil det i psykologiske eksperimenter være passende at sætte prior odds for H_1 relativt til H_o til omkring 1:10, svarende til $P(H_o) \approx 0.90$.

Vil man disse problemer til livs, ligger den oplagte løsning lige for: sænk signifikansniveauet. Benjamin et al. (2018) har foreslået en generel sænkning af det konventionelle 5% signifikansniveau til 0.5%, altså en markant højere overlægger for at kunne erklære et resultat 'statistisk signifikant'. Dette forslag er dog ikke uden problemer (se de afsluttende kommentarer nedenfor).

2.2 Sandsynligheden for H_o, givet data.

Den klassiske p-værdi angiver sandsynligheden for de observerede data, eller mere ekstreme data ($D+$), givet H_o, dvs. $P(D+ \mid H_o)$. Det er en udbredt misforståelse blandt empirikere, at en lav p-værdi kan tages som udtryk for, at der er lav sandsynlighedsmæssig evidens for H_o, og dermed høj sandsynlighedsmæssig evidens for H_1 (jf. Gigerenzer, 2004; Wasserstein and Lazar, 2016; Harvey, 2017). Uformelt fortolker vi ofte en lav p-værdi som udtryk for, at $P(H_o \mid D)$ er lav. En medvirkende årsag til denne misforståelse kan givetvis spores tilbage til Fisher's fortolkning af en lav p-værdi (som nævnt ovenfor): "*Either an exceptionally rare chance has occurred, or the theory is not true.*"

Men dette leder til '*the fallacy of the transposed conditional*'. Når vi beregner p-værdien, er det betinget på fordelingen af teststatistikken under H_o, dvs. vi betinger på, at H_o er sand. Dermed fortæller p-værdien ingenting om hverken den ubetingede sandsynlighed $P(H_o)$, eller den betingede sandsynlighed $P(H_o \mid D)$, og dermed heller ingenting om sandsynligheden for H_1.

Vi kan vha. Bayes formel beregne $P(H_o \mid D)$ og $P(H_1 \mid D) = 1 - P(H_o \mid D)$, som en funktion af á priori sandsynligheden $P(H_o)$, likelihood funktionen for de observerede data, samt en á priori fordeling for parameteren af interesse. Udtrykt ved Bayes-faktoren (BF) fås (jf. Engsted, 2019):

$$P(H_o \mid D) = \frac{\text{BF} \cdot P(H_o)}{1 + [P(H_o)(\text{BF} - 1)]}. \tag{2}$$

Hvis vi anvender den Bayes-faktor, der for en given t-statistik giver maksimal sandsynlighedsmæssig evidens imod H_o, dvs. $\text{BF} = \exp(-\frac{1}{2}t^2)$, jf. Berger and Sellke (1987), fås for $t = 1.96$ (svarende til en p-værdi på 0.05) BF$= \exp(-\frac{1}{2}1.96^2) = 0.146$. Med neutrale á priori odds for hypoteserne ($P(H_o) = 0.50$), fås hermed fra formel (2): $P(H_o \mid D) = 0.128$, altså en markant højere sandsynlighed end p-værdien.

Hvis $P(H_o \mid D)$ skal ækvivalere p-værdien (0.05), skal $P(H_o) = 0.265$

og $P(H_1) = 0.735$.[3] Med andre ord: Når vi forkaster H_o på et 5% signifikansniveau, opererer vi implicit med den á priori opfattelse, at H_1 er væsentlig mere sandsynlig end H_o. Bemærk, at dette gælder selv for den Bayes-faktor, der giver maksimal sandsynlighedsmæssig evidens imod H_o. For mindre ekstreme - mere rimelige - Bayes-faktorer, indebærer en forkastelse på 5% niveau med et klassisk t-test, en endnu højere á priori sandsynlighedsopfattelse af H_1, dvs. $P(H_o) < 0.265$ og $P(H_1) > 0.735$. Det er mit indtryk, at de færreste empirikere er klar over den ret skæve implicitte prior, der ligger gemt i en forkastelse på det traditionelle 5% signifikansniveau med det klassiske test. Og $P(H_o) < 0.50$ er under alle omstændigheder ikke forenelig med Berger and Sellke's (1987), Harvey's (2017) og Benjamin et al.'s (2018) anbefaling om, at $P(H_o) > 0.50$, som beskrevet ovenfor.

Ofte bliver valget af et lavt signifikansniveau (eksempelvis 5%) begrundet som følger: H_o udtrykker vores '*maintained hyothesis*' eller '*working hypothesis*' som kræver meget stærk evidens imod sig i stikprøven før vi er villige til at forkaste H_o. Dette kræver, at hvis vi forkaster, skal sandsynligheden for at begå en fejl være lav, svarende til en lav Type-I fejlsandsynlighed. Men denne argumentation fører til et paradoks, for som vi netop har set, indebærer en forkastelse på 5% niveau en forhåndsformodning om, at H_o er *mindre* sandsynlig end H_1, altså svarende til at det i virkeligheden er H_1 som er vores working hypothesis.

Startz (2014) illustrerer paradokset med et konkret eksempel og konkluderer: "*We usually think that our standards for significance are chosen precisely to point in the direction of the null unless we have strong evidence to the contrary. But as this example illustrates, our usual standards do not accomplish that goal. In other words, in this example the p-values we usually regard as providing strong evidence against the null and in favor of the alternative do not in fact provide such evidence unless the econometrician already leaned strongly toward the alternative.*" (Startz, 2014, p.141).

2.3 Multiple tests.

Signifikansniveauet (α) i det klassiske test er udtryk for Type-I fejlsandsynligheden i *repeated sampling* ved udførelse af ét test. Sættes $\alpha = 0.05$, og under antagelse af at H_o er sand, vil - på tværs af mange gentagelser af testet på nye stikprøver - 5% af testene fejlagtigt afvise H_o.

I de fleste forskningsprojekter udføres der dog ikke kun ét test i hver

[3] Fra formel (2) fås $P(H_o) = \frac{P(H_o|D)}{P(H_o|D)(1-\mathrm{BF})+\mathrm{BF}} = \frac{0.05}{0.05(1-0.146)+0.146} = 0.265$.

stikprøve. I en given stikprøve foretages der typisk mange forskellige tests. Der afprøves forskellige (kombinationer af) forklarende variable i forskellige modeller og for forskellige delperioder, der laves forskellige misspecifikationstests, etc. Der udføres med andre ord multiple tests. Hvis der vælges et 5% signifikansniveau i hvert enkelt test, vil det samlede signifikansniveau for hele batteriet af tests være højere end 5%.

Hvis der udføres to uafhængige tests, hver på 5% niveau, bliver det samlede signifikansniveau $1 - (1 - 0.05)^2 = 0.0975$ (9.75%). Hvis det samlede signifikansniveau skal være 5%, kræver det et signifikansniveau i hvert test på 2.53%. Udføres der 10 tests, skal signifikansniveauet i hvert test være 0.51%, svarende stort set til at dividere det ønskede samlede signifikansniveau (5%) med antallet af tests (Bonferroni-korrektionen). I praksis vil de udførte tests ikke være uafhængige, hvilket komplicerer korrektionen. Der er i øvrigt et væld af spidsfindigheder involveret i korrektion for multiple tests, se Harvey et al. (2020) for en god gennemgang af de forskellige metoder.[4]

I empiriske studier indenfor samfundsvidenskab korrigeres der sjældent for multiple tests. Der anvendes typisk et 5% signifikansniveau i hvert enkelt test. Dermed er sandsynligheden for, at der undervejs i sekvensen af tests begås Type-I fejl (væsentligt) højere end 5%, hvorved sandsynligheden for *false discovery* bliver høj (jf. formel (1)).

3. Kulturændring i vores empiriske praksis.

Der er mange paradokser og indbyggede modsætninger i det klassiske hypotesetestsetup. De konventionelle signifikansniveauer (1%, 5%, 10%) er arbitrære og bliver brugt på samme måde, uanset om det er H_o eller H_1, der udtrykker vores arbejdshypotese (den model vi undersøger), og valget af signifikansniveau har en indbygget skjult - og uerkendt - forhåndsopfattelse af hypoteserne, der langt fra er neutral eller objektiv. De klassiske hypotesetest bygger endvidere på et *repeated sampling* setup, der i teorien kræver en meget detaljeret og forudbestemt stikprøveplan, og som sjældent er dækkende for den type passivt observerede, ikke-eksperimentelle data samfundsforskere

[4]Harvey et al. (2020) diskuterer også afvejningen mellem *false discoveries* og *missed discoveries*, hvor sidstnævnte henviser til det problem, der opstår, hvis man med henblik på at reducere den samlede Type-I fejl, sænker signifikansniveauet til et sådant niveau, at man kommer til at begå for mange Type-II fejl, altså undlade at forkaste en forkert hypotese.

arbejder med (Engsted, 2020). Dertil kommer, at lærebogsfremstillinger af teorien for hypotesetest oftest er en hybrid af Neyman-Pearson metoden (valg mellem to hypoteser) og Fisher's metode baseret på p-værdien (forkastelse eller ikke-forkastelse af H_o), på trods af, at de to tilgange faktisk er uforenelige (Hubbard and Bayarri, 2003).

Vi kan godt gå rundt og lade som om - eller forsøge at overbevise os selv og hinanden om - at vi i vores empiriske forskning ikke er dikteret af rigide konventionelle signifikansniveauer. Men praksis viser noget andet, jf. indledningen til afsnit 2: Signifikans på 5% niveau fungerer stadig som en de facto overlægger for at kunne erklære et videnskabeligt resultat! Denne praksis fører på den ene side - som argumenteret ovenfor - til 'påvisning' af for mange effekter og sammenhænge, der ikke reelt eksisterer (eller som er ubetydelige), men indebærer samtidig en underprioritering af forhold, der burde spille en langt større rolle i empiriske analyser, så som økonomisk signifikans (fremfor statistisk signifikans) og modelusikkerhed (fremfor stikprøveusikkerhed).

Som nævnt ovenfor har Benjamin et al. (2018) foreslået en generel sænkning af det konventionelle 5% signifikansniveau til 0.5% som et 'quick fix' til at rydde ud i de mange *false discoveries*. Men spørgsmålet er, om udskiftning af én arbitrær konvention med en anden er vejen frem. Problemerne med begrebet statistisk signifikans stikker langt dybere, og flere statistikere og empirikere har på det seneste givet endnu mere markante anbefalinger: "Moving to a world beyond "p<0.05"" (Wasserstein et al., 2019), "Abandon statistical significance" (McShane et al., 2019), "Retire statistical significance" (Amrhein et al., 2019). Under alle omstændigheder bør replikationskrisen få os til seriøst at diskutere, om der er behov for en kulturændring i vores empiriske praksis.

Referencer.

Amrhein, V., S. Greenland, B. McShane + 800 signatories (2019): Retire statistical significance. *Nature* 567, 305-307.

Andrews, I., and M. Kasy (2019): Identification of and control for publication bias. *American Economic Review* 109, 2766-2794.

Benjamin, D.J., Berger, J.O., et al. (2018): Redefine statistical significance. *Nature Human Behavior* 2(1), 6-10.

Berger, J.O., and T. Sellke (1987): Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association* 82, 112-139.

Engsted, T. (2019): Bayesianske hypotesetests. I: *Symposium i Anvendt Statistik 2019* (red. Peter Linde), Københavns Universitet og Det Nationale Forskningscenter for Arbejdsmiljø.

Engsted, T. (2020): Likelihoodprincippet og den klassiske p-værdi. I: *Symposium i Anvendt Statistik 2020* (red. Peter Linde), Aarhus Universitet og Det Nationale Forskningscenter for Arbejdsmiljø.

Engsted, T., and J.W. Schneider (2022): Non-experimental data, hypothesis testing, and the likelihood principle: A social science perspective. Preliminary and incomplete working paper, Aarhus University.

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

Fisher, R.A. (1959): *Statistical Methods and Scientific Inference* (2nd ed.). Oliver and Boyd Ltd., Edinburgh.

Gigerenzer, G. (2004): Mindless statistics. *Journal of Socio-Economics* 33, 587-606.

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

Harvey, C.R., Y. Liu, and A. Saretto (2020): An evaluation of alternative multiple testing methods for finance applications. *Review of Asset Pricing Studies* 10, 199-248.

Hubbard, R., and M.J. Bayarri (2003): Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *American Statistician* 57, 171-178.

Ioannidis, J.P.A., T.D. Stanley, and H. Doucouliagos (2017): The power of bias in economics research. *Economic Journal* 127, F236-F265.

McShane, B.B., D.G. Gal, A. Gelman, C. Robert, and J.L. Tackett (2019): Abandon statistical significance. *American Statistician* 73, 235-245.

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

Storey, J.D. (2003): The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics* 31, 2013-2035.

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

Wasserstein, L.R., A.L. Schirm, and N. Lazar (2019): Editorial: Moving to a world beyond "p<0.05". *American Statistician* 73, 1-19.

How to construct Individual preference scales

Gorm Gabrielsen, Professor emeritus, Copenhagen Business School

The present work is a methodological study to investigate the possibly construction of individual preference scales based on paired comparisons. The data material is consumer preference of pork in Denmark.

A set of 56 consumers tasted four samples of meat as paired comparison. All of the six possible pairs of four samples of meat were served to each consumer. In each comparison of two meat samples, the preference was scored on a 15 cm visual analog scale. The consumer was asked to mark on a line, which sample he or she preferred. A mark right between the two meat samples means that they equally well liked or disliked the samples. The closer the mark was to one of the sides the more was this sample preferred. The meat was served to the consumers in a sensory laboratory (ISO 8589).

After a 10 minutes break the consumer scored the same four samples of meat on a 15 cm visual analog scale. This makes it possible to compare the results appearing from the use of hedonic scales to the results appearing from the application of paired comparisons.
Using the two methodologies results at the aggregated level basically in the same differences in preferences, but it seems that the consumers differentiates the samples more in the paired comparisons. In addition, the method of paired comparisons makes the individual differences visible. This shows a heterogeneity among consumers, which question the concept of "mean consumer".

For the consumer to rate a meat sample into a hedonic scale often requires use of preliminary sorting. The preliminary sorting may be related to individual values - mothers cooking - or it may be related to family values or to cultural values – sometimes named "meta preferences". The choice between the use of hedonic scale or paired comparison therefore involves the question of which kind of preferences is of interest in a specific situation.

En simpel data-dreven Bayes faktor

By Klaus Rostgaard

Danish Cancer Society Research Center, Danish Cancer Society, Copenhagen

klar@cancer.dk

Abstract

Traditionel frekventistisk testning fører i praksis for ofte til afvisning af nulhypotesen (falsk positive) bl.a. fordi der ikke tages hensyn til sandsynligheden for alternativet. Bayesiansk metodik gør, men ofte er antagelserne for at gøre det enten subjektive eller utroværdige. Der findes en approximativ formel for Bayes faktor som er monoton i likelihood-ratioen (og dermed p-værdien), som er simpel, og som kan anvendes bare man kender dimensionen af alternativ-hypotesen, og den relevante teststørrelse eller p-værdi. Denne Bayes faktor afhænger også af en parameter, der udtrykker balancen i information mellem prior og data. Vi vil give et bud på hvordan den vælges. Standard-udgaven af denne Bayes faktor bliver en kontinuert udgave af Akaike informationskriteriet til modelselektion.

Note: Dette indlæg bygger på et indsendt paper, derfor er der ingen detaljer her.

Introduktion

Den Bayesianske læreproces for en parameter θ består i at opdatere en a priori tæthed $p(\theta)$ med hvad vi lærer fra data X i form af en likelihood funktion $L(\theta;X)$ til en ny tæthed $p(\theta|X) \propto L(\theta;X) \times p(\theta)$. Denne Bayesianske læreproces har en pendant for fordelingen af sandsynligheden for at en given model/hypotese M_k er den rigtige i et endeligt univers af modeller $M_0,\ldots M_K$. Pendanten virker ved at a priori model odds, $Pr(M_k)/Pr(M_0)$ opdateres i lyset af data X v.h.a. den såkaldte Bayes faktor (BF_{k0}) til at opnå a posteriori model odds $Pr(M_k|X)/Pr(M_0|X)=BF_{k0} \times Pr(M_k)/Pr(M_0)$. Så Bayesiansk hypotesetestning går ud på at bruge denne formel til at udtale sig om hvor sandsynlige de forskellige betragtede modeller er, efter at have set data. $BF_{k0} = Pr(X|M_k)/Pr(X|M_0)$ og er dermed en ratio af prediktiv performance på data X, og noget der ligner en likelihood-ratio.

Det er klart ud fra disse formler at a priori model odds er noget som læseren selv kan vælge efter subjektivt forgodtbefindende, mens man ved præsentation af resultater nok bør sætte a priori model odds til 1, svarende til at denne prior er det eneste mulige kompromis mellem to opponenter der har preference for hver sin model/hypotese, og i øvrigt svarer til at evidens fra data X er alt vi vil bringe i spil. Derfor er BF det eneste der er vigtigt at forholde sig til.

En universel Bayes faktor

Indenfor epidemiologi (og mange andre fag) er Cox, Poisson og logistisk regression nærmest eneherskende til parametrisk modellering, og de relevante tests er næsten altid på formen H_0: $\theta=0$ versus H_1: $\theta\neq 0$, hvor estimatet for θ under H_0 antages asymptotisk normalfordelt centreret i 0, $N_d(0,V)$, og dermed at et likelihood -ratio test for H_1 versus H_0 er χ^2-fordelt med d frihedsgrader, hvor d er dimensionen af θ.

Når vi prøver at udbygge dette aymptotiske maskineri på Bayesiansk vis, og kræver monotoni i p-værdien (og formler der generaliserer naturligt fra én til flere dimensioner – i én dimension er p-værdi monotoni altid sikker) lander man lynhurtigt i én bestemt formel for Bayes factor:

$$BF_{10} = \psi^{d/2}LR^{1-\psi} \Rightarrow \log(BF_{10}) = d/2\ \log(\psi) + (1-\psi)/2\ \ \chi^2$$

hvor LR er likelihood-ratioen for testet, χ^2 er forskellen i deviance mellem de tilsvarende modeller og $\psi=\lambda/(1+\lambda)$ hvor λ udtrykker en ratio af information mellem prior og data, idet vi har antaget prioren under H_1 til at være $N_d(0, \lambda^{-1}V)$.

Hvordan vælges λ?

BF_{10} er unimodal som funktion af ψ. Hvis ψ gøres meget lille svarer det til at smøre sandsynlighedsmassen tyndt ud og dermed have meget lidt sandsynlighedsmasse i omegnen af maximum-likelihood estimatet for θ (og derfor foretrække H_0). Hvis ψ gøres meget stor svarer det til at koncentrere sandsynlighedsmassen omkring 0, og vi lærer ikke rigtigt noget af data (og foretrækker H_0). BF_{10} maximeres (størst evidens for alternativet H_1) for $\psi=d/\chi^2$. Asymptotisk er χ^2 en Wald-teststørrelse, så for at få λ til at skrumpe i det rigtige tempo når informationen i data (V^{-1}) vokser vælger vi ψ ud fra $\lambda= d/\chi^2$.

Bayes faktor er som matematisk konstruktion symmetrisk i de to modeller/hypoteser. Det vil derfor være naturligt at kræve at Bayes faktor skal være monotont voksende i χ^2 og at den for passende små værdier af χ^2 skal være <1 (altså understøtte H_0 mere end H_1). Det kan man opnå ved at vælge $\lambda= \min(d/\chi^2, \lambda_{max})$. Hvis man fx vælger λ_{max} = 0.255 fører det til en Bayes faktor der har præcis de samme preferencer som hvis man anvender Akaikes informationskriterium til at vælge modeller, hvor vi foretrækker den simplere model (H_0) hvis forbedringen i deviance per dimension er mindre end 2.

Hvordan bliver denne inferens i praksis?

Hvis vi tænker på likelihood-ratio testet og BF som funktioner af data X, der inddeler parameterrummet i regioner hvor vi hhv. accepterer og afviser H_0, så bliver formen på disse regioner den samme, men grænsen mellem de to regioner ligger et andet sted,

således at BF generelt vil udpege en større region som kompatibel med H_0. Og dermed får vi færre falsk positive testresultater. Som nævnt opfører standard-versionen af BF sig som om vi lavede modelselektion ud fra Akaikes informationskriterium. Ved at vælge λ tilpas lille (svarende til en meget svag prior der indeholder information svarende til én observation eller ét udfald) kan vi får BF til at opføre sig som det Bayesianske informationskriterium, som er konsistent, men som på den korte bane favoriserer H_0 temmelig meget. Mulighederne er mange.

Bemærk endelig, at hvis vi estimerer, fastlægger data altid i det lange løb a posteriori fordelingen, og prioren er i den forstand derfor ikke kritisk ved estimation. Hvis vi derimod tester eller predikterer slipper vi aldrig af med konsekvenserne af vores valg af prior.

Finansministeriets anbefalede diskonteringsrente er en skat på vores børn og børnebørn[1]

Jesper Jespersen, Roskilde Universitet

Det nittende århundrede strakte 'den finansielle kalkulation' til det yderste som kriterium for, om såvel private som offentlige aktiviteter var anbefalelsesværdige. Hele livsførelsen blev herved gjort til en parodi på en bogholders mareridt: 'kan det betale sig?' Den samme selvdestruktive finansielle overvejelse hersker over alle tilværelsens aspekter: Vi ødelægger landskabs skønheder, fordi de glæder ved naturen, der ikke lader sig omsætte i penge, ikke har nogen økonomisk værdi. Vi ville om nødvendigt være parat til at slukke månen og stjernerne, fordi de ikke giver dividende.
John Maynard Keynes (1933)

Samfundsøkonomisk tab: Drivhusgas og arbejdsløshed

Hvad har emission af drivhusgasser til atmosfæren og de timer, der mistes ved ufrivillig arbejdsløshed, til fælles?

De er irreversible. Emission af drivhusgas vil forøge den globale temperatur for altid. Ligesom ufrivillig arbejdsløshed indebærer et tab af produktion (og velfærd) for altid.

Den politiske ambition burde derfor være at nedbringe emissionen af drivhusgasser i et hastigt tempo frem mod år 2030 og helt af afvikle emissionen inden år 2050; for herved at bidrage til, at temperaturstigningen ophører i løbet af det 21. århundrede.

Samtidigt hermed befinder dansk samfundsøkonomi (på lige fod med de fleste andre vestlige økonomier) sig i den situation, at der i de kommende år vil være en betydelig ufrivillig arbejdsløshed, som der ikke er en målrettet plan for at mindske. Tværtimod vil der være en betydelig risiko for, at europæisk økonomi bliver paralyseret af krav om budgetbalance og nedbringelse af den offentlige gæld. Altså en gentagelse af forløbet efter finanskrisen, hvor ikke mindst vedtagelsen af Budgetloven og EU's finanspagt stillede sig hindrende for en genopretning af balancen i dansk økonomi.

[1] Opdateret og revideret kapitel 2.2 fra bogen 'Kriseøkonomi og Klimagæld', Forlaget Jensen og Dalgaard, 2021

Denne mangel på samfundsøkonomisk planlægning er paradoksal, idet den øgede arbejdsløshed, jo giver en unik mulighed for at anvende ledige ressourcer (arbejdskraft og 'gravkøer') til at gennemføre den politisk ønskede omstilling af dansk økonomi.

Tillad mig at give blot et par historiske eksempler: Lillebæltsbroen i 1930erne og etablering af Storebæltsforbindelsen i første halvdel af 1990erne. Perioder med ekstraordinær høj arbejdsløshed. Jeg kender ikke én fagøkonom, der i dag udtrykker andet end fuld anerkendelse for gennemførelsen af disse ganske store anlægsprojekter.

Behovet for gennemførelse af en 'grøn omstilling' er lige så åbenbart. Det vil nemlig stille krav om betydelige investeringer i: vind-, sol- og jordenergi, omstilling af industri, af transport, af landbrugsproduktion og isolering og opvarmning af boligmassen. Skal denne omstilling lykkes, vil det kræve, at der hver år iværksættes ekstra private og offentlige investeringer i størrelsesordenen af en halv Storebæltsforbindelse, dvs. 20-30 mia. kr. Set i forhold til det danske bruttonationalprodukt på ca. 2.400 mia.kr. er det dog ikke noget overvældende beløb – og slet ikke, når de nødvendige produktionsressourcer står ledige!

Benspænd for den *samfunds*økonomiske rationalitet

Realiseringen af denne omstilling vil dog kræve: for det første at Budgetloven ændres, og for det andet at finansministeriets vejledning vedrørende vurdering og anbefaling af offentlige investeringer – herunder miljø- og klimainvesteringer – ændres. Den skal bringes i samklang med den samfundsøkonomiske virkelighed: karakteriseret ved stigende arbejdsløshed, lav realrente og beskeden BNP-vækst de næste mange år.

Budgetloven opererer med et krav om strukturel balance på den offentlige sektors saldo. Det vil sige at underskud op til 3 pct. af BNP kun kan accepteres i helt ekstraordinære krisesituationer, som f.eks. Covid19-krisen 2020-22; men efterfølgende er kravet balance på budgettet. Dog har et bredt flertal i folketinget netop bevilliget en undtagelse fra Budgetloven i form af et ekstraordinært underskud på ca. 10 mia. kr., der dog er øremærket til øgede forsvarsudgifter. Men så løftede finansministeriet også pegefingeren overfor andre ensidige udgiftsstigninger. (ja, et underskud på ½ pct. af BNP er tilladt; men sigtet skal være balance). Det betyder, at et løft i det grønne investeringsniveau kun kan gennemføres, hvis der spares andre steder på de offentlige budgetter, eller den længe ventede CO2-afgift vedtages og provenuet øremærkes til klimainvesteringer. https://altandetlige.dk/blog/jesperjespersen/vil-krisen-skabe-opbrud-makrooekonomisk-teori-799

Finansministeriet anbefaler en høj diskonteringsrenten

Men budgetloven, som folketinget kunne ændre med et pennestrøg, er ikke den eneste blokering for den grønne omstilling. Lige så blokerende er finansministeriets krav om et *real*afkast på alle offentlige investeringer på 4 pct. (, der ligeledes bygger på den teoretiske misforståelse, at samfundsøkonomien kan analyseres som værende i en langsigtet tilstand af fuld ressourceudnyttelse: generel ligevægt (sic!), og at den 'rigtige' samfundsøkonomiske rente kan fastsættes som værende lig med den forventede reale vækst plus et (betydeligt) risikotillæg. Begge antagelser fører til en diskonteringsrente, der resulterer i en massiv underinvestering i den offentlige sektor navnlig i perioder med langvarig arbejdsløshed, overopsparing i den private sektor og lav væk

Boks: Tekst og tabel fra Finansministeriet, 2021:

Opdatering og sænkelse af den samfundsøkonomiske diskonteringsrente

Finansministeriet har besluttet at opdatere og sænke den anbefalede samfundsøkonomiske diskonteringsrente fra 4 pct. til 3,5 pct. i år 0-35, fra 3 pct. til 2,5 pct. i år 36-70 og fra 2 pct. til 1,5 pct. efter 70 år. Den gældende samfundsøkonomiske diskonteringsrente er illustreret i tabellen nedenfor.

Tabel 1
Den opdaterede samfundsøkonomiske reale diskonteringsrente

	0-35 år	36-70 år	>70år
Real diskonteringsrente	**3,5 pct.**	**2,5 pct.**	**1,5 pct.**
Risikofri realrente	2 pct.	1,75 pct.	1,5 pct.
Risikopræmie (ikke-diversificerbar risiko)	1,5 pct.	0,75 pct.	0 pct.
Memo-post			
Real statsobligationsrente i 2025-fremskrivning	0,5 pct.	2 pct.	2 pct.

Anm.: Den faldende profil for den samfundsøkonomiske diskonteringsrente anvendes på den måde, eksempelvis for et projekt der løber over 75 år, at den del af projektets omkostninger og gevinster, der realiseres i løbet af de første 35 år, diskonteres med en rente på 3,5 procent pr. år, mens gevinster og omkostninger, der ligger mellem år 36 og år 70 diskonteres med en rente på 2,5 pct., og for år 71-75 anvendes 1,5 pct..

Ovenstående tabel stammer fra Finansministeriets vejledning til offentlige myndigheder vedrørende valg af diskonteringsrente ved gennemførelse af samfundsøkonomiske cost-benefit analyser:Der kan stilles mange undrende spørgsmål til

denne tabel, der anbefaler en real diskonteringsrente på 3½ pct. i 35 år efterfulgt af en uforståelig aftagende tidsprofil[2].

Lad mig begynde med den uforståelige tidsprofil: at risikotillægget reduceres jo længere ude i fremtiden indtægter og udgifter ligger? Et er vel sikkert, at jo længere ude i fremtiden en begivenhed ligger desto større er usikkerheden – eller hvad? Men i lyset af den høje diskonteringsrente på 3½ pct. p.a. og dernæst 2½ pct. p.a. de første 35/70 år, så kunne renten såmænd efterfølgende sættes til nul eller til 5 pct. i de følgende år. Det ville under alle omstændigheder kun have en minimal indflydelse på nutidsværdien af investeringen, da de fremtidige indtægter og udgifter allerede efter 70 år er nedskrevet til mindre end 10 pct.[3]

'Den samfundsøkonomiske diskonteringsrente': Prisen på fremtiden

Diskonteringsrenten er et udtryk for det ekstra beløb, som kræves af investor for at en fremadrettet investering har en positiv nutidsværdi. Er renten 4 pct. p.a. så kræves det, at en investering i dag på 100 mill. kr. om ét år skal betales tilbage med 104 mill. kr.

I tabellen nedenfor ses der på en investering, der løber over 35 år, hvor finansministeriets rentekrav er 4 pct. p.a. Det betyder, at investeres der 100 mill. kr. i en vindmølle, der producerer elektricitet, skal der om 35 år betales 400 mill. kr. tilbage med rentes rente. Det er derfor fuldt berettiget at kalde 'renten': enten for 'prisen på fremtiden' eller måske endnu mere korrekt for en 'skat på vore børn og børnebørn'.

Problemet er blot, at sådan fremstilles det ikke i den offentlige debat. Her anses en betydelig positiv forrentning nærmest som et indiskutabelt krav – prøv at spørge din bank! Det giver måske mening, så længe vi taler om et privatøkonomisk lån, hvor renten til en vis grad er at ligne med en forsikringspræmie. Hvis låntager går fallit, er pengene (delvis) mistet. Men i en samfundsøkonomisk

[2] *Den samfundsøkonomiske diskonteringsrente*, Finansministeriet, 12. november 2018. Her var anbefalingen 4 pct. – men efter hårdt pres ikke mindst efter Eldrup-udvalgets redegørelse, hvor en sådan høj diskonteringsrente satte et spørgsmålstegn ved den samfundsøkonomiske rentabilitet af omstillingen til el-biler. Der blev efterfølgende udarbejdet en anbefaling af at mindske diskonteringsrenten til 3½ pct. og et supplerende notat, https://fm.dk/media/18371/dokumentationsnotat-for-den-samfundsoekonomiske-diskonteringsrente_7-januar-2021.pdf. Hvorfor 'risikopræmien' falder efter 35 år, henstår i det uvisse; men det har stort set ingen betydning, idet nutidsværdien af de beløb, der indgår i CB-analysen efter 35 år med en realrente på 3½ pct., er diminutiv.

[3] Det er netop det problem, som de bevilligende myndigheder slås med i forbindelse med dekommissionering af atomreaktoren på Risø. Da den blev planlagt for mere end 50 år siden synede udgiften til afvikling af reaktoren diminutiv, da man dengang benyttede en endnu højere real diskonteringsfaktor. Men dyrt blev det den dag afviklingen blev påbegyndt for ca. 10 år siden og er endnu ikke afsluttet.

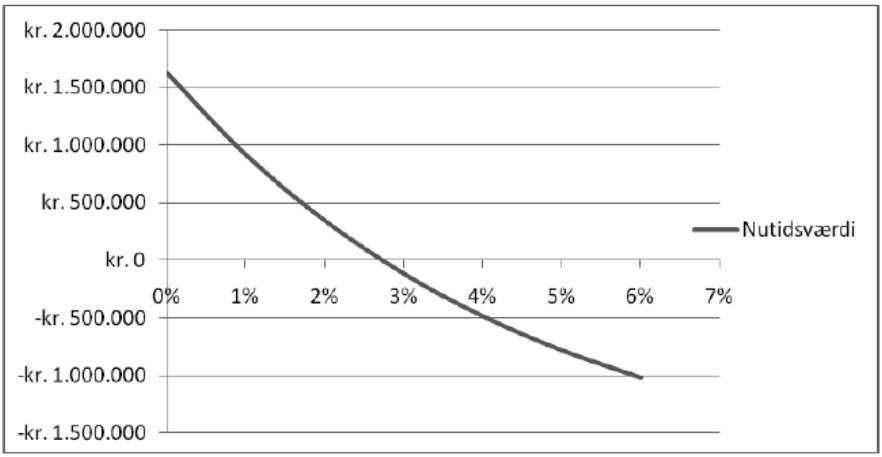
kalkulation, hvor det er samfundet eller rettere 'slægtsgården', der skal vedligeholdes, og det er staten, der står som låntager og garant, er problemstillingen snarere omvendt. Investeringer i bæredygtig udvikling gør slægtsgården mere robust over for bl.a. klimatiske omskiftelser og mindsker herved den samfundsmæssige risiko, som er knyttet til andre private eller offentlige investeringsprojekter, der rækker ud i fremtiden.

Som illustreret i nedenstående figur er et højt afkastkrav til fremtidssikring af samfundsøkonomien en sikker måde at forarme kommende generationer på. Jo højere rente desto færre investeringer i en bæredygtig fremtid. Det fører simpelthen til underinvesteringer ikke alene i miljøbeskyttelse, klimaforbedring, men også i undervisning, forskning og infrastruktur.

Kravet om den høje diskonteringsrente synes i dag helt løsrevet fra den samfundsøkonomiske virkelighed. Alene af den grund, at staten kan låne til en *real*rente, der længe ligefrem har været negativ! Så politikerne burde udfordre finansministeriet: Hvor kommer dog argumentationen om så en høje en realrente fra?

Jeg bliver i hvert fald læseren svar skyldig. Men efter at have læst vejledningen står det indtryk tilbage, at finansministeriet muligvis forveksler samfundsøkonomi med privatøkonomi.

Figur 1. Nutidsværdi og diskonteringsrente



Figur 1. Kalkulationsrentens betydning ved investering i havvindmøller. Alle beløb i 1000 kr.

Kilde: Den samfundsøkonomiske kalkulationsrente – fakta og etik, Concito, notat, februar, 2011 (så dette er ikke nogen ny erkendelse!)

Diskonteringsrente, bæredygtig udvikling og kommende generationer

Det andet mindst lige så tungtvejende argument imod en høj diskonteringsrente er af overvejende etisk karakter og negligeres derfor ofte i den økonomiske litteratur. Enhver renteberegning er som nævnt ovenfor reelt en afvejning mellem nutid og fremtid. Et krav om en positiv rente/afkast er således en form for byrde/beskatning af fremtidige generationer. En positiv diskonteringsrente betyder, at den nuværende generation kræver af kommende generationer et over tiden stigende beløb som tilbagebetaling af et givet lånebeløb. Fastsættes renten til 4 pct., betyder det med rentes rente, at om blot 17 år skal der betales et beløb tilbage, der er vokset til det dobbelte og om 35 år til det firdobbelte! Det kan vist uden overdrivelse kaldes en betydelig beskatning af kommende generationers velfærd, jeg fristes til at sige livsvilkår, som således er indeholdt i finansministeriets vejledning.

Jeg ville have svært ved at se mine børn og børnebørn i øjnene, hvis jeg stillede et krav om, at de penge, som de har brug for at låne af mig i dag, skulle betales tilbage med rente og rentes rente. Hvorfor gælder det samme etiske princip ikke for offentlige investeringer? De er jo nødvendige fællesskabsudgifter til at sikre, at der er en velfungerende samfund(søkonomi), når børn (og børnebørn) når min alder. Det er en 'slægtsgård', som vi har fået ansvaret for at forvalte og sikre vedligeholdelsen af. Vi har kun fået den til låns af vore forældre og med (moralsk) pligt til at give den videre i ordentlig stand til den næste generation.

Det kan desværre med en betydelig ret hævdes af Gretha Thunberg m.fl., at den nuværende generation (født i efterkrigstiden) på en række områder har misrøgtet 'slægtsgården'. Den dyne af CO_2, som er blevet lagt udover kloden og som stadig forøges med rivende hast år for år, er der ganske enkelt ikke blevet betalt for. Det i mange andre sammenhænge gode princip, at 'forureneren betaler for oprydningen', burde vel også gælde her?

Men hvordan ryddes der op i atmosfæren? Teknologien er endnu ikke udviklet hertil. Det er derimod den teknologi, der kan skabe og sikre en endog særdeles effektiv produktion af vedvarende energi. Det vil dog kræve massive 'grønne' investeringer, som den nuværende (specielt ældre) generation burde forestå og betale, for herigennem at bidrage til at miljøgælden vokser langsommere og helt ophører med at vokse efter år 2050, se Jespersen, *Vækstøkonomi på vildspor*, 2019!

En anden måde at betale denne miljøgæld af på kunne være at renoncere på kravet om et positivt afkast på disse investeringer. Ja, det moralsk korrekte ville vel

ligefrem være at give et tilskud til disse investeringer i form af en negativ rente. Det ville fremskynde omstillingen til bæredygtig økonomi. Hvorved den samfundsmæssige risiko for et kollaps mindskes, hvilket som nævnt vil gøre private investeringer mindre risikobehæftede til fordel for kommende generationer.

Et krav om en positiv realrente (der ligger udover den realøkonomiske vækst[4]) på investeringer i en bæredygtig fremtid er slet og ret en beskatning af kommende generationer. Hvorfor skal personer, der i dag sparer op egentlig have et på forhånd fastlagt tillæg i form af et renteafkast i en samfundsøkonomi, der mildt sagt har en usikker fremtid? En usikkerhed der ydermere forstærkes, hvis der ikke investeres tilstrækkeligt i bæredygtig udvikling.

Hvilket leder frem til det samfundsøkonomiske paradoks – eller måske rettere dilemma: Jo højere kravet om afkast er (fra de grådige pensionsopsparerne), desto mindre bliver der investeret i bæredygtig omstilling, hvilket forøger risikoen for et samfundsmæssigt sammenbrud, at væksten går i stå og måske ligefrem bliver negativ.

Det samfundsetiske perspektiv tilsiger således, at fremtiden burde indgå med mindst samme vægt som nutiden ved beregningen af, om 'det kan betale sig' at gennemføre 'grønne' investeringer. Det vil sige, at der skulle benyttes en diskonteringsrente på nul. Ja, måske ligefrem en negativ rente, hvis den omhandlede investering nedsætter risikoen for et sådant sammenbrud i fremtiden. Investering i en grøn og bæredygtig udvikling er jo at sammenligne med en slags lynafleder, hvis værdi de færreste betvivler!

Konklusion: Slægtsgården forfalder, hvis der ikke investeres i fremtiden

Hvis det perspektiv, at samfundet kan lignes med en slægtsgård, accepteres, så burde enhver generation som en selvfølge føle sig forpligtet af en 'samfundskontrakt', hvori vedligeholdelse af 'slægtsgården' indgår som en selvfølge. Man kunne kalde det en moralsk forpligtelse til at sikre, at slægtsgården overdrages i god stand til den næste generation. Det er – kort fortalt – den neoliberale økonomis forbandelse, at den har individuel optimering som sit teoretiske grundlag. Her er værdigrundlaget som bekendt individuel optimering af nutidsværdien: Det er udgifter, som *jeg* har nytte af, der skal optimeres – jfr. bl. a. Keynes' kritik i det

[4] Dette forbehold skal naturligvis vurderes i lyset af den usikkerhed der i dag hersker om der overhovedet er nogen rimelig forventning om en realøkonomisk vækst pr. capita i dansk økonomi – herom strides de lærde.

indledende citat af sin samtids økonomer: det såkaldte '*Treasury* View on Money and Finance'.

Lægges denne neoklassiske, privatøkonomiske og individualistiske optimeringsteori til grund for vurderingen af offentlige investeringer, så bliver 'slægtsgården' misligholdt; 'for det kan jo ikke betale sig'. Så jeg har en betydelig forståelse (og sympati) for den anklage Greta Thunberg's rettede mod de 'gamle mænd' på mødet i U.N.'s Climate Action Summit in New York City, september 2019 med ordene: *'How dare you?*'

https://www.npr.org/2019/09/23/763452863/transcript-greta-thunbergs-speech-at-the-u-n-climate-action-summit

Så lad mig slutte, hvor jeg begyndte: at finansministeriets anbefaling af at benytte en diskonteringsrente på tidligere 4 pct. nu 3½ pct. p.a. indebærer, at der spilles unødigt hasard med vore børns og børnebørns livsvilkår, for det fører til underinvestering i 'slægtsgården'.

Litteratur:

Finansministeriet, *Den samfundsøkonomiske diskonteringsrente*, november 2018 (opdateret i 2021)

Jespersen, J. *Miljøøkonomi*, Djøfs Forlag, 1998 (kan stadig fås som e-bog)

Jespersen, J. *Vækstøkonomi på Vildspor*, København: Jensen & Dalgaard, 2019
(Essays om økonomi, politik og virkelighed igennem 20 år)
https://jensenogdalgaard.dk/shop/non-fiktion/326-jd.html

Jespersen, J. *Kriseøkonomi og Klimagæld: hvor skal pengene komme fra?* København: Jensen & Dalgaard, 2021

Keynes, John Maynard "National Self-Sufficiency" *The Yale Review*, Vol. 22, no. 4 (June 1933), pp. 755-769. https://jmaynardkeynes.ucc.ie/national-self-sufficiency.html

Fertility and Promotions - Academic careers of economists over 40 years in Denmark

Anne Sophie S. Lassen, CBS, and Ria Ivandic, Oxford/LSE

The arrival of children often takes place early in the research career. For many researchers, the life-changing and demanding event of becoming a parent overlap with the period the researcher has to show high levels of research productivity with the aim of gaining tenure, or an equivalent promotion to a professorship. It is important to understand how having children contributes to the 'the leaky pipeline' of women and which factors mitigate or exacerbate gender gaps in attrition following parenthood. Our project asks how fertility decisions, and specifically their timing, can affect women's academic careers in Economics. We will study this question with administrative data from Denmark covering the universe of individuals entering PhD programs in economics over a 40-year period, linked with high-quality bibliometric data from Scopus. To understand how the arrival of children affects productivity and how this maps on to women's probability of a successful career in academia, we first estimate the unconditional child penalty of women and men economists on career progression (i.e. attrition and promotion). We then move on to estimate the child penalty of women and men economists on productivity outcomes, namely number of publications and quality of publications. While being aware that productivity is endogenous to fertility, we will estimate whether the child penalty in career progression persists once we control for research productivity. High-frequency data on productivity and family outcomes will allow us to compare whether the child penalty varies across the timing of birth in different stages of the academic career, allowing us to draw insights into how the timing of children interacts with pressure on early career productivity and to what extent this disadvantages women. Finally, we will explore under which circumstances gender gaps are mitigated and focus on two channels: first, the current position held by the researcher and the structure of promotion and characteristics of their workplace and second, the individual circumstances depending on their partner's career path.

Om måling af det umålelige

Anvendelse af personlighedstest til etablering af studiegrupper

Mogens Dilling-Hansen

Department of Economics and Business Economics

Aarhus Universitet

Mail: dilling@econ.au.dk

Resumé

I denne analyse er der fokus på personlighedstests og deres anvendelse til inddeling af nye studerende i studiegrupper. Antallet af metoder til bestemmelse af persontype er lige så mangfoldig som antallet af forslag til personlige karakteristika, der har indflydelse på en persons effektivitet i en given opgave-/team-sammenhæng. Den videnskabelige dokumentation af gevinsten ved at anvende personlighedstests er ikke entydig, men der er konsensus om, at det er vigtigt at bestykke teams med forskellige persontyper. I den empiriske del undersøges det metodiske grundlag for inddeling af nye erhvervsøkonomiske studerende i studiegrupper, og der anvendes data for 2018 og 2022. Der påvises en vis usikkerhed om måleskalaens reliabilitet og specielt er den eksterne validiteten et problem, så længe de latente strukturer ikke kan vurderes i forhold deres sande værdier.

1 Indledning

Vi ved det godt. Somme tider klarer en person en opgave bedre end andre ville have gjort det, og selv om der er mange ressourcemæssige årsager til det (evner, flid...), så er der stadig noget tilbage, der kan beskrives som type af personlighed. At vi er forskellige er nok ikke så interessant i økonomisk forstand, men det er til gengæld interessant at undersøge eventuelle sammenhænge mellem personlighedstype og evne til at klare en opgave.

Formålet med denne analyse er at undersøge, om sammensætning af studiegrupper på universitetsniveau efter typen af personlighed er en god ide. Populationen, der undersøges, er erhvervsøkonomistuderende ved Aarhus Universitet; de sidste 5 år er studerende inddelt i studiegrupper med henblik på at reducere frafaldet, og tankegangen er den, at studiegrupper sammensat på den "rigtige måde" vil ikke alene øge læringen for den enkelte studerende, men også skabe positive afledte sociale og læringsmæssige effekter, der samlet ikke alene vil være til glæde for den enkelte studerende men også for hele studiet.

Opbygningen af analysen er som følger. Kapitel 2 indeholder en oversigt over personlighedstests på det danske marked med særlig fokus på måling af de opstillede typer. Kapitel 3 er en præsentation af det konkrete gruppedannelsesprojekt og data for hhv. 2028 og 2022. Analyserne præsenteres i kapitel 4, hvorefter der rundes af i kapitel 5 rundes af med en diskussion af de fundne resultater.

2 Om personlighedstest

Anvendelse af OpenAI's chatbot ChatGPT giver følgende beskrivelse af en personlighedstest:

"*En personlighedstest er et redskab, der anvendes til at måle og beskrive en persons personlighed. Der findes mange forskellige typer af personlighedstest, men de fleste af dem har til formål at indsamle oplysninger om en persons følelser, tanker, holdninger, værdier og adfærdsmønstre. Formålet med personlighedstestene kan være at hjælpe individer med at forstå deres egne personlighedstræk og at hjælpe organisationer med at vælge de bedst egnede medarbejdere til bestemte job. Personlighedstestene kan også bruges til at forudsige hvordan en person vil reagere i bestemte situationer eller hvordan de vil fungere i en gruppe.*" (kilde: ChatGPT).

Mao. der etables et link mellem en persons type og evnen at bestride et job.

Selve opgaven, at klassificere personer efter typen af personlighed, er måske mindre interessant, medens det at forskellige typer af personlighed i forskellig grad kan have indflydelse på hvordan personen klarer et bestemt job, er af langt større relevans.

Der er mange typer af personlighedstest, og alene en simpel dansk google-søgning på "personlighedstest" giver 103.000 hits. En lang række af de opstillede tests er grundlæggende god underholdning uden reel substans, hvilket understreges af den gruppe af udbydere, der anvender personlighedstyper (check fx https://www.disc-profil.dk/blog/personlighedstyper). Andre tests tilbydes gratis af organisationer, der er aktive i jobsøgningsprocessen, se fx KRIFA og JOBINDEX, medens den store gruppe af personlighedstests er en del af et samlet kommercielt tilbud, hvor en persons type matches med et givet job. Her skal omtales tre typer, MBTI, DISC og Belbin Test.

MBTI er interessant, fordi den opstillede procedure til etablering af personlighedstyper var en af de første (Isabel Briggs Myers udviklede konceptet i perioden 1920 til 1940) og fordi der var en direkte reference til Jungs psykologiske forskning (se Jung (1976) og MBTI): Baseret på Jung bliver alle personer i MBTI beskrevet ved deres personlighedstype i en binær form

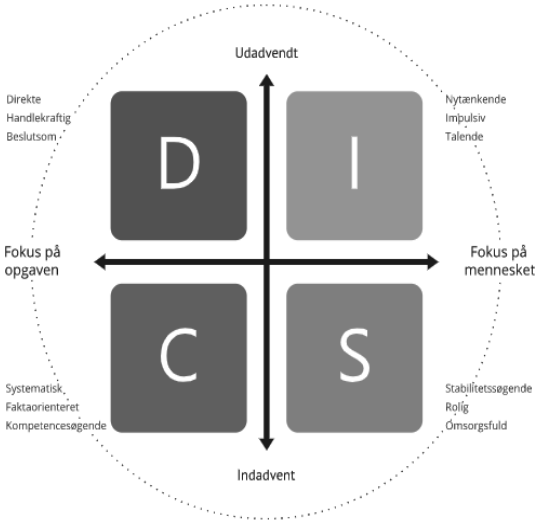
i) Fokus på omverdenen -> Ekstrovert eller introvert
ii) Tilgang til vurdering af information -> Analytisk eller intuitiv
iii) Tilgang til beslutninger -> Baseret på ræsonnementer eller følelse
iv) Tilgang til beslutninger -> Baseret på vurdering eller opfattelse

De fire dimensioner kan udfoldes til 16 (2^4) forskellige persontyper, der hver især er særligt velegnede til forskellige jobtyper.

DISC profileringen af personer (se DISC) er på tilsvarende måde et veletableret kommercielt produkt, der også er baseret på den psykologiske forskning og som også inddeler personer i farver – argumentet er igen, at din personlighedstype (her beskrevet via farven) bestemmer din adfærd og dermed også din egnethed til forskellige typer af jobs (se DISC).

Opdelingen i typer vha. farverne rød, gul, grøn og blå er i sig selv interessant; men nok så interessant er det, at personens personlighed kædes sammen med to generelle dimensioner: Personens generelle ageren (introvert-ekstrovert) og personens tilgang til løsning af en opgave (fokus på jobbet eller mennesket).

Figur 1 DISC profilens opbygning – dominans, indflydelse, stabilitet og kompetencer



Kilde: https://www.disc-profil.dk/blog/hvad-er-disc-profil

Det sidste kommercielle redskab (Belbin test) til at etablere personlighedstyper er valgt, fordi det eksplicit søger at etablere et direkte link mellem den enkelte person og rollen i et team – tankegangen er, at jo bedre match, der er mellem en personlighedstype og krav til et givet job, jo bedre vil et team af personer være i stand til at løse en given opgave. En særlig styrke ved denne metode er, at ophavsmanden (Meredith Belbin) har rod i den engelske akademiske verden, og som følge deraf, er der publiceret en række videnskabelige artikler om metoden. Dels har Belbin selv publiceret bøger om metoden (se fx Belbin (2010)) og dels er der flere artikler, der søger at etablere videnskabeligt belæg for sammenhæng mellem personlighedstyper og effektivitet af teams.

Det er naturligvis en anelse bekymrende, at konklusionerne af de videnskabelige bidrag er lidt forskellige; i Batenburg et. al. (2013) findes på basis af et studium af 24 teams IKKE nogen korrelation mellem sammensætning og performance af et team, og det er på linje med andre tilsvarende studier: En stor andel af studierne kan ikke påvise nogen sammenhæng, medens resten påviser ”some evidence”.

Swailes et. al. (2002) rejser desuden det grundlæggende spørgsmål, om der faktisk måles korrekt, når en person klassificeres – baseret på analyser af Cronbach's Alpha konkluderes det blandt andet, at der er behov for yderligere analyser, før validiteten kan bevises.

Figur 2 Belbin test – etablering af 9 grundlæggende teamroller

Belbins 9 teamroller overblik

Teamrolle		Beskrivelse af rollen og bidraget til teamarbejdet	Tilladelige svagheder, der kan følge med rollen
Idémand		Begavet, kreativ og idérig. Ser vanskelige problemstillinger fra nye vinkler.	Kan være svag i sin kommunikation med andre. Glemsom og ikke praktisk anlagt. Kritikfølsom.
Kontaktskaber		Udadvendt, entusiastisk, nysgerrig og meddelsom. Undersøger muligheder. Skaber kontakter	Flygtig. Taber let interessen, når den første entusiasme har lagt sig. Taler meget.
Koordinator		Moden, selvsikker og tillidsfuld. Kan prioritere. Klargør mål og frembringer beslutninger. Har øje for andres talenter.	Kan have en tendens til at manipulere og være imperiebygger. Ikke nødvendigvis den mest vidende i teamet.
Opstarter		Dynamisk, højt gearet og rastløs. Udfordrer og skaber pres, finder veje uden om forhindringer.	Kan have et heftigt temperament. Er utålmodig, påståelig og stædig. Kan virke provokerende.
Analysator		Analytisk, nøgtern og objektiv. Præcis dømmekraft. Ser alle rationelle aspekter af en sag.	Opfattes ofte som meget direkte, kritisk og skeptisk. Noget træg og ikke så inspirerende for andre.
Formidler		Socialt orienteret, udadvendt og skarpt iagttagende. Sensitiv, diplomatisk og fleksibel. God lytter. Undgår gnidninger og skaber et godt klima.	Kan være ubeslutsom og usikker i afgørende situationer. Kan være overfølsom.
Organisator		Disciplineret, pålidelig og loyal. Effektiv i gennemførende faser. Realistisk og praktisk.	Noget ufleksibel. Reagerer langsomt overfor nye muligheder og er langsom i sin tilpasning.
Afslutter		Omhyggelig og samvittighedsfuld. Leder efter fejl og forglemmelser. Perfektionistisk, vedholdende og præcis.	Kan have en tendens til at bekymre sig unødvendigt. Emsig samt bange for at begå fejl. Utilbøjelig til at delegere.
Specialist		Bidrager med specialviden og tekniske færdigheder. Stærk fagligt engagement og selvtillid. Meget koncentreret om sine mål og opgaver.	Tendens til at isolere sig og være uinteresseret i andre mennesker. Vogter sit område og bidrager snævert inden for dette.

Kilde: https://potential.dk/belbins-9-teamroller/?gclid=EAIaIQobChMI_ICckNnv6gIVhKZ3Ch2wHwGZEAAYASAAEgJAtfD_BwE

De tre metoder, MBTI, DISC og Belbin, er alle eksempler på, hvorledes en persons type og en givet opgave bør matches, og deres valg af metode er ifølge dem selv, baseret på årelang erfaring ... og baseret på videnskabelig evidens.

Der er naturligvis grænser for, hvor stor grad af transparens, der kan forventes af kommercielle virksomheder, mht. etablering af personlighedstyper; en nødvendig men ikke tilstrækkelig betingelse for, at måling af typer er reliabel (Cronbach's alpha) og valid (stabilitet over tid) er dog, at denne dokumentation er tilgængelig. Et væsentligt problem med at teste reliabilitet og validitet på denne måde er påpeget af Cho et. al. (2015): Højere reliabilitet kan opnås (ved at fjerne variable), men det er ofte på bekostning af validiteten (oprindelig inklusion af en variabel er baseret på validitetsargumenter).

I det efterfølgende analyseres etablering af personlighedstyper med henblik på at bestemme arbejdsgrupper for nystartede erhvervsøkonomi-studerende i 2018 og 2022.

3 Måling af personlighedstyper – etablering af data

De erhvervsøkonomiske uddannelser ved Aarhus Universitet producerer cand. merc.'er, og et kendetegn ved disse studerende er, at de har let ved at finde et job og at jobbet er godt lønnet. Bagsiden af medaljen er, at et relativt stort antal studerende dropper ud allerede i løbet af det første studieår. Det kan der være gode grunde til, men både for den studerende, der dropper ud "for tidligt" og for Aarhus Universitet er omkostningerne høje, og derfor er der siden 2018 dannet studiegrupper, således studerende udover at de samles på øvelseshold (ca. 30 studerende) også – på frivillig basis - tilbydes en studiegruppe bestående af 4-5 studerende.

Metoden til inddeling i studiegrupper er inspireret af de gængse metoder til at udføre personlighedstests, herunder de tre metoder præsenteret i kapitel 2. Den grundlæggende tankegang er, at arbejde i studiegrupper vil øge studieegnetheden for den enkelte studerende, at studiegrupper på 4-5 studerende vil øge både det sociale netværk og øge læringen, at de studerende ikke er ens udrustet med henblik på at studere, og at studerende med forskellige kompetencer/personligheder vil skabe den bedste dynamik i en studiegruppe.

Oplysninger til etablering af persontype er spørgeskemabaseret, dvs. baseret på selvrapportering, og det er principielt tale om totaltælling af hele årgangen af nye studerende. Ca. 1% af de nye

studerende, der formelt er indskrevet som studerende den 20. august, har ikke ønsket at deltage og der er i 2018 og 2022 informationer for hhv. 778 og 731 studerende.

Metoden til etablering af studiegrupper er dels inspireret af metoder til bestemmelse af personlighedstyper (dvs. med fokus på studentens evne/tilgang til at samarbejde) og dels målrettet til den arbejdssituation, som en studiegruppe medfører. Der måles i fire dimensioner:

i)	Dirigenten	–> leder, effektiv, målrettet
ii)	Redaktøren	-> disciplineret, velstruktureret, detaljeorienteret
iii)	Idealisten	-> nytænkende, velovervejet, kreativ
iv)	Facilitatoren	-> fleksibel, positiv, social

Tankegangen er, at en studiegruppe vil fungere bedst, hvis der både er en til at lede gruppen, en til at få arbejdet til at ske, en til at lave analyserne og en til at skabe den gode stemning; men hvor det understreges, at ingen af kompetencerne kan stå alene. Hver dimension etableres på baggrund af 6 holdningsspørgsmål, og der anvendes en 5-punkts Likert-skalering; den efterfølgende gruppeinddeling sker ved, at der udtages en studerende fra toppen af hver af de 4 dimensioner (principielt simpelt tilfældigt, men med hensyn til en fornuftig kønsmæssig sammenhæng). Som eksempel på de stillede spørgsmål anvendes 6 spørgsmål til etablering af den studerendes "dirigent-egenskaber":

Item 1:
Jeg er god til at motivere mine gruppemedlemmer til at bidrage til arbejdet.
Item 2:
Selvom jeg som udgangspunkt er interesseret i alles synspunkter, kan jeg hurtigt tage et standpunkt, når en beslutning skal træffes.
Item 3:
Jeg er god til at fordele arbejdsopgaverne mellem gruppens medlemmer med det formål at optimere processen.
Item 4:
Jeg har en tendens til at blive utålmodig, når de andre i gruppen ikke arbejder i samme tempo som mig. / Jeg kan have en tendens til at blive dominerende, hvis der er behov for at få noget gennemført.
Item5:
Jeg er ikke bange for at sige min mening, selvom jeg måske er i undertal i den pågældende situation.
Item 6:
Jeg ser mig selv som en form for ledertype og kan træffe den endelige beslutning, hvis det bliver nødvendigt.

Allerede efter første runde (2018) var der behov for revision af et at de stillede spørgsmål (item 4), så i de efterfølgende reliabilitets- og validitetsanalyser er dette spørgsmål udeladt; det kan dog bemærkes at de fundne resultater er robuste og er ikke påvirket af, at item 4 er udeladt.

4 Analyse af reliabilitet og validitet af faktorer til gruppedannelse

For hvert individ, der deltager i undersøgelsen, bestemmes personlighedstypen ved at beregne en score for hver af de fire dimensioner/roller: Dirigent, redaktør, idealist og facilitator. I denne analyse af målingens validitet og reliabilitet fokuseres der på "dirigent"-rollen; analyserne er også gennemført for de øvrige analyser og resultaterne er helt på linje med nedenstående.

Til vurdering af de opstillede måleskalaer for de 4 personlighedstyper anvendes sammenligning af gennemsnit og beregning af Cronbach's α.

$$\text{Cronbach's } \alpha = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum_{i=1}^{k} \sigma_y^2}{\sigma_x^2}\right)$$

hvor k er antallet af items (her 5), tælleren i variansudtrykket er variansen for hver item, medens nævnerens varians er for summen af de 5 items. Cronbach's α angiver hvor meget de udvalgte items korrelerer med hinanden og ligger i intervallet 0 og 1, og selv om målet ikke er et formelt test på reliabilitet, så antages værdier på 0.7 og derover at være acceptabelt.

Tabel 1 "Dirigent-egenskaber" for 2018 og 2022; gennemsnit og Cronbach's α

	2018 Gennemsnit	Cronbach's Alpha	2022 Gennemsnit	Cronbach's Alpha
Item 1	3.79	0.62	3.68**	0.68
Item 2	3.89	0.59	3.83	0.66
Item 3	3.93	0.61	3.82**	0.64
Item 5	4.06	0.63	3.92**	0.67
Item 6	3.66	0.51	3.49**	0.57
Samlet	778	0.65	731	0.70

*Noter: De beregnede gennemsnit for 2022 er sammenlignet med et 99% konfidensinterval beregnet for 2018 – en signifikant afvigelse er anført med **. Cronbach's α for de 5 items er anført i rækken "Samlet"; de partielle α'er ud for hver item er Cronbach's α, givet det konkrete item udelades.*

De studerendes vurdering af egne personlige egenskaber er givet årsagen til at alle spørgsmål har et gennemsnit over 3 (skalering er 1-5), og det ses at der i 4 ud af 5 tilfælde er en signifikant ændring i målingen af den gennemsnitlige "dirigent-egenskab" og i alle tilfælde er der tale om et signifikant fald. Udgangspunktet for måling af denne ordinale information over tid er, at niveauet bør være nogenlunde uændret, så dette fald er generelt problematisk, medmindre der er andre faktorer, der kan forklare faldet. Det faglige niveau (målt ved adgangsgivende karaktergennemsnit) er stort set uændret, ligesom evt. corona-effekter ikke bør invalidere analysen (2018 og 2022 ligger uden for den periode, hvor der var formelle restriktioner).

Vurderes reliabiliteten af måleskalaen for "dirigent-egenskaber" ud fra Cronbach's α er der på tilsvarende vis grund til en vis bekymring. Det kan endvidere anføres, at der for de øvrige dimensioner er tilsvarende problematiske værdier af Cronbach's α: 0.75 for redaktør, 0.55 for idealist og 0.54 for facilitator; men for alle dimensioner er der tale om en stigende reliabilitet fra 2018 til 2022 ... baseret på Cronbach's α.

En nødvendig men ikke tilstrækkelig betingelse for valide måleskalaer er at de er reliable. Selv om resultaterne er lidt bekymrende, så er reliabilitet lettere at teste end validitet, fordi udover kravet om at der måles konsistent, så skal det også være "det rigtige", der måles. Det er et meget ambitiøst mål at måle latente personlige karakteristika, specielt når der kigges på måling af personlige egenskaber, som ikke er objektive og som i bedste fald måles på en ordinal skala; dermed er der grænser for, hvor formelt validitet kan vurderes. Ud fra ønsket om ekstern validitet, som sikrer at resultater kan generaliseres, kan der trods alt opstilles en række betingelser, der skal være opfyldte: Udover at der måles reliabelt, så må bestemmelse af personlighedstyper også i en vis grad udvise stabilitet over tiden. Tabel 1 viser, at selv om den anvendte måleskala måske er acceptabel, så er det alligevel bekymrende, at den gennemsnitlige vurdering for de bagvedliggende items, der bestemmer personens karakteristika, falder signifikant over tiden.

Tabel 2 Eksplorativ faktoranalyse på "Dirigent-egenskaber" for 2018 og 2022

	2018			2022		
	Loadings	Communality	KMO MSA	Loadings	Communality	KMO MSA
Item 1	0.41	0.17	0.74	0.43	0.18	0.77
Item 2	0.53	0.28	0.73	0.53	0.28	0.76
Item 3	0.45	0.20	0.70	0.57	0.33	0.74
Item 5	0.44	0.19	0.72	0.47	0.22	0.77
Item 6	0.79	0.62	0.68	0.80	0.65	0.70
Eigenvalue	2.10 / 29.3%			2.26 / 33.1%		

Noter: Modellerne er estimeret vha. Maximum Likelihood og der er anvendt CFA. Begge modeller har kun egenværdi over 1 og er statistisk signifikante (Bartlett's test of Sphericity). Tommelfingerregler siger, at MSA bør være mindst 0.7, medens anbefalingen for loadings ligger i intervallet 0.4-0.7.

Vurdering af faktoranalysen på "dirigent-spørgsmålene" præsenteret i tabel 2 viser en statistisk signifikant model og pæn stabilitet i factor loadings over tid, men resultaterne illustrerer også svagheden ved korrelationsanalyser med mange observationer forstået på den måde, at den fælles latente faktor reelt ikke kan forklare de manifeste variable. Baseret på faktoranalysen kan det derfor ikke afvises, at måleskalaen er valid; men kun i tilfælde af, at den "sande værdi" af den latente variabel kendes, vil det endeligt kunne afgøres om måleskalaen er valid.

5 Afrunding

Det er på mange måder interessant at betragte den store mængde af personlighedstests, først og fremmest fordi for de kommercielle versioner anfører, at et match mellem personens type og et givet job/team vil øge effektiviteten. Den videnskabelige litteratur på områder finder dog ikke samme entydige sammenhæng.

For gruppedannelsesprojektet for nye erhvervsøkonomistuderende ved Aarhus Universitet har der været positive tilbagemeldinger fra de studerende om tiltaget, men fordi der ikke er anvendt et eksperimentelt design, så er det ikke muligt at vurdere effekten på frafaldet. Givet er det, at metoden til inddeling i studiegrupper følger de gængse metoder, og at der også er en vis usikkerhed omkring, hvilke personlige karakteristika, der måles.

Måske skal personlighedstests bare ses som et øjebliksbillede af en persons opfattelse af sig selv ... og så er det måske helt andre kriterier, der skal anvendes til at inddele studerende i studiegrupper.

Referencer

Arrow, K. J. (1951), "Alternative Approaches to the Theory of Choice in Risk-Taking Situations", *Econometrica*, vol. 19, no. 4: 404-437.

Batenburg, R, Wv Walbeek & Wid Maur (2013), Belbin role diversity and team performance: is there a relationship?, *Journal og Management development, vol 32, no 8, 2013, pp 901-913.*

Belbin, R. M. (2010) "Management Teams", 3rd edition, Routledge, London

Cho, E. & S. Kim (2015), Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, vol 18, no 2, pp. 207-230.

Jung, C.G. (1976), "Psychological Types (The Collected Works of C. G. Jung, Vol. 6)", oprindelig publiseret 1921, International Kindle Paperwhite.

Swailes, S & McIntyre-Bhatty, T. (2002), The "Belbin" team role inventory: reinterpreting reliability estimates", *Journal of Managerial Psychology*, research note, vol 17, no 6, 2002, pp 529-536.

***Anvendte internetbaserede kilder**: (alle websites er sidst refereret december 2022)*

Belbin test, https://potential.dk/belbins-9-teamroller/?gclid=EAIaIQobChMI_ICckNnv6gIVhKZ3Ch2wHwGZEAAYASAAEgJAtfD_BwE, Potential, Kokkedal

ChatGPT, https://openai.com/, OpenAI

DISC, https://www.disc-profil.dk/blog/hvad-er-disc-profil, DISC-Profil, Herlev

JobIndex, https://www.jobindex.dk/persontypetest?lang=da, JobIndex, Valby

Krifa, https://krifa.dk/udvikling/test-din-persontype, Krifa

MBTI, https://www.myersbriggs.org/my-mbti-personality-type/mbti-basics/, Myers & Briggs Foundation

Stiger antal smittede af kønssygdommene lige meget?

Anders Milhøj

anders.milhoj@econ.ku.dk

Abstrakt

I de senere år er antal personer smittede med klamydia, gonorre og syfilis steget nærmest eksplosivt, efter at smittetrykket var nær ved nul først i halvfemserne. I dette indlæg undersøges, om smitten udvikler sig ensartet for mænd og kvinder, og for hhv. heteroseksuelle og homoseksuelle mænd.

Data

Data er hentet fra Seruminstituttets hjemmeside, hvor månedlige antal indberettede tilfælde af syfilis og gonorre, begge anmeldelsespligtige. og antal positive klamydiaprøver løbende offentliggøres. Data offentliggøres også opdelt på smitte i indland/udland og aldersgrupper, men i denne fremstilling benyttes kun opdelingen i køn, idet antal smittede for mænd opdeles i kvinde-mand og mand-mand som smittevej for mænd, der betegnes hhv. hetero- og homoseksuelle mænd, selvom disse ord er misvisende for biseksuelle. Der benyttes altså tre grupper for gonore og syfilis, mens antal klamydiatilfælde kun opdeles efter køn. Det kunne være interessant at benytte de øvrige opdelinger, men det udsættes.

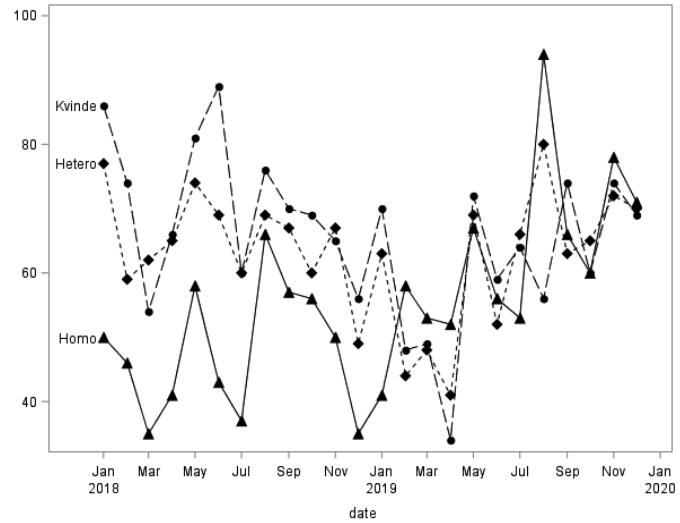
Niveauerne stiger voldsomt, måske endda eksponentielt, og sæsoneffekten og den irregulære komponent er tydeligt multiplikativ, så der skal transformeres. I stedet for en logaritmetransformation vælges en kvadratrodstransformation, da antal syfilistilfælde er så små, at en del måneder er helt uden smittede. I teorien kan fordelingen af antal tilfælde pr. måned siges at være Poissonfordelt, og da middelværdien i denne fordeling er lig med variansen, er det naturligt at benytte en kvadratrodstransformation.

Grafisk præsentation af tidsrækkerne

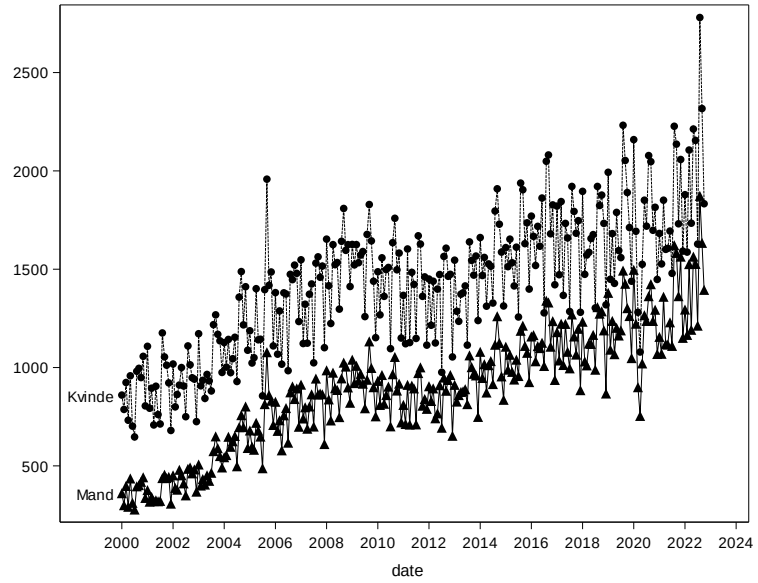
Forsøges de tre tidsrækker for de månedlige antal gonoretilfælde for hhv. kvinder, heteroseksuelle mænd og homoseksuelle mænd plottet i samme plot, udvikler de tre tidsrækker sig så ens, at plottet bliver uden informationsværdi. Niveauet stiger fra under 10 pr. måned for alle tre grupper til omkring 120 de seneste måneder i 2022. Der er en klar sæsonvariation, der dog ikke er helt ens i de tre tidsrækker, se Figur 1 der viser rækkerne i de to kalenderår 2018 og 2019. Figur 2 viser udviklingen i de månedlige antal positive klamydiatilfælde for mænd og kvinder. Det er markant, at der er flere positive klamydiaprøver blandt kvinder end blandt mænd i hele perioden, og at niveauet stiger voldsomt til ca. 2000 månedlige tilfælde blandt kvinder og 1300 tilfælde blandt mænd.

Der er en sæsonvariation med toppe i januar og august, se Figur 3, der viser rækkerne i de to kalenderår før coronaen.

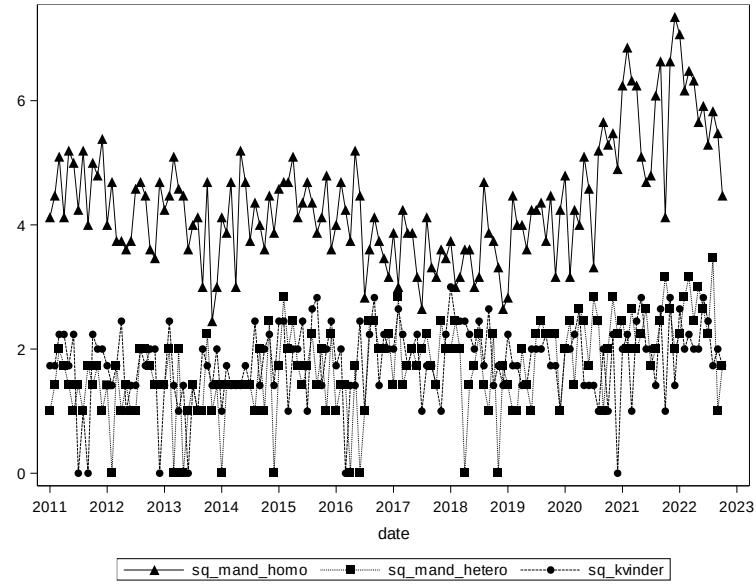
Figur 1 Antal gonorretilfælde for hhv. kvinder, mænd smittet af mænd og mænd smittet af kvinder i 2018 og 2019



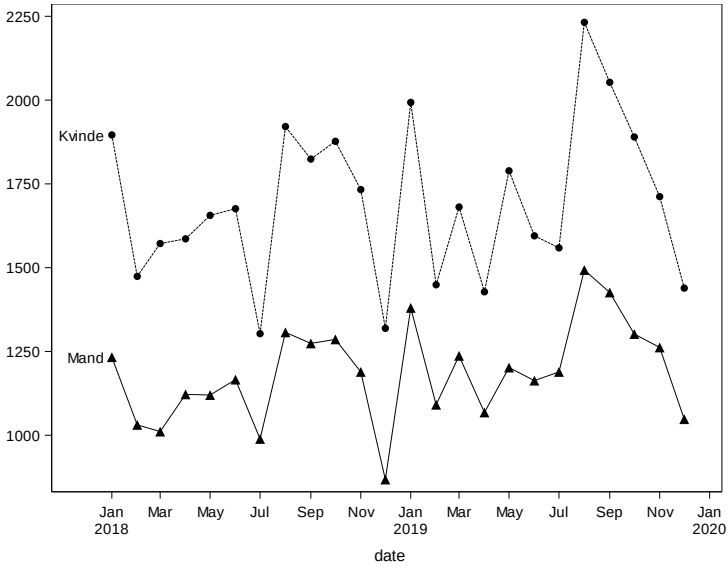
Figur 2 Antal klamydiatilfælde for hhv. kvinder og mænd



Figur 3 Antal klamydiatilfælde for hhv. kvinder og mænd i 2018 og 2019



Figur 4 Kvadratroden af antal syfilistilfælde for hhv. kvinder, mænd smittet af mænd og mænd smittet af kvinder



Der er få syfilistilfælde især i de tidlige år i data, og der er endda mange måneder helt uden tilfælde. Figur 4 viser kvadratroden af de månedlige antal tilfælde for hhv. kvinder, heteroseksuelle mænd og homoseksuelle mænd. Det ses, at for kvinder og heteroseksuelle mænd er mange observationer nul, og kun få, især i de sidste måneder, er større end fire, dvs. med kvadratrod to. Niveauet for homoseksuelle mænd er klart højere end for de to andre tidsrækker. Der kan ikke spores sæsoneffekter, da de i givet fald er slørede i den store variation i Poissonfordelinger med lav intensitet.

I de følgende afsnit analyseres tidsrækkerne for de tre kønssygdomme for hver sygdom for sig, og til sidst samles de i alt otte tidsrækker i en samlet model.

Metode

Da der er sæson i de fleste af tidsrækkerne, tages udgangspunkt i sæson ARIMA modellerne, dvs. i den generelle sæson ARIMA(p, d, q) × ARIMA$_S$(P, D, Q) model

$$(1 - \varphi_1 B - .. - \varphi_p B^p)\,(1 - \Phi_1 B^{12} - .. - \Phi_P B^{12P})(1 - B)^d(1 - B^1)^D X_t =$$

$$(1 - \theta_1 B - .. - \theta_q B^q)\,(1 - \Theta_1 B^S - .. - \Theta_Q B^{12Q})\varepsilon_t$$

I dette udtryk betegner B backshift (lag) operatoren og restledsprocessen, ε_t, er hvid støj. Selve tidsrækken X_t kan være flerdimensional, når alle parametrene er matricer.

Desuden benyttes dynamisk faktoranalyse, hvor en n-dimensional tidsrække, X_t, modelleres ved hjælp af en k-dimensional latent tidsrække λ_t ved

$$X_t = B\lambda_t + \varepsilon_t$$

Dette udtryk kan suppleres med observerede forklarende variable og et middelværdiled. Her betegner B en n × k dimensional matrix med faktorloadings og ε_t er en k-dimensional vektor af ukorrelerede variable.

Den k-dimensionale latente tidsrække λ_t består af tidsrækker, der fx. kan modelleres ved sæson ARIMA modeller.

I denne estimeres disse modeller ved brug af Proc SSM i SAS. Parametrene estimeres ved Kalman filteret, der minimerer kvadratsummen af forudsigelsesfejlen ved prædiktion en måned frem. Disse forudsigelser plottes i flere af de følgende figurer for at demonstrere modellernes fit i praksis.

Analyse af klamydiatilfælde for to køn

For de to tidsrækker for antal positive laboratorieprøver for hhv. mænd og kvinder benyttes en sæson ARIMA model som en enkelt latent faktor, og da antallet for kvinder er langt højere end for mænd, skal der også anvendes en niveauforskydning. Da det er tydeligt på grafen, at forskellen indsnævres mellem de to rækker indsnævres benyttes også en lineær trend for forskellen.

Den latente faktor følger en ARIMA(0,1,1)×ARIMA(1,0,0) model

$$(1 - \Phi_1 B^{12})\,(1 - B)\lambda_t = (1 - \theta_1 B)\,\varepsilon_t$$

Modellen bliver for X_1, mænd og X_2, kvinder

$X_{1t} = \lambda_t + w_{1t}$

$X_{2t} = \mu + \beta \times t + \lambda_t + w_{2t}$

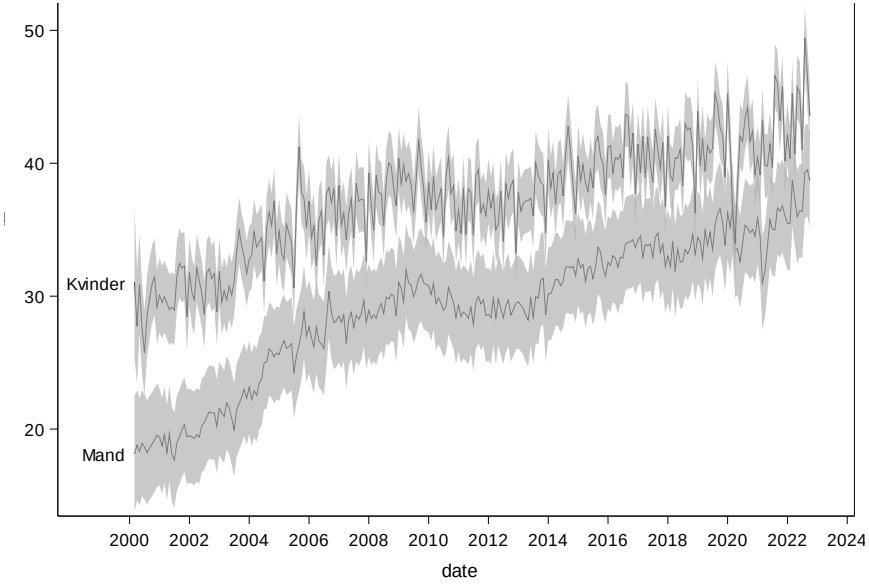
I denne model estimeres parametrene til

$\theta_1 = 0.81$, $\Phi_1 = 0.66$, $\mu = 16.83$, $\beta = -0.00046$

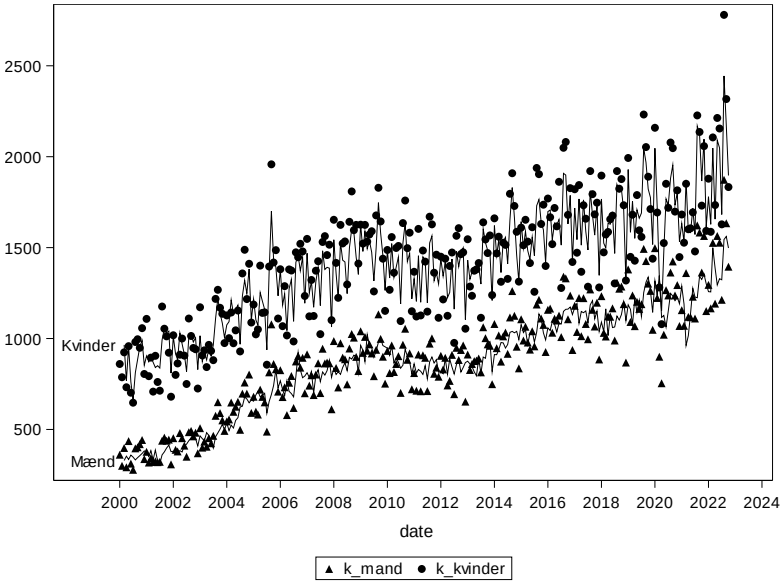
Alle parametrene er stærkt signifikante. Den lave numeriske værdi af det estimerede β trods t-værdien –15.32, skyldes, at tidsbetegnelsen er i SAS' standard, dvs. antal dage siden 1/1 – 1960, der jo er voldsomt store tal.

Modellens fit ses af Figur 5, der viser forecasts med sikkerhedsgrænser for de to tidsrækker. Effekten af det signifikante β ses tydeligt på, at afstanden mellem de to bånd indsnævres. Figur 6 viser disse forecasts samt de originale data efter en transformation tilbage til de oprindelige antal.

Figur 5 Forudsigelse med sikkerhedsgrænser for antal klamydiatilfælde for kvinder hhv. mænd



Figur 6 Forudsigelse for antal klamydiatilfælde for kvinder hhv. mænd



Analyse af de tre tidsrækker for syfilis

Observationerne for heteroseksuelle mænd og for kvinder ligger på samme niveau gennem alle årene, så det er naturligt at forsøge sig med samme model for de to tidsrækker. Efter en række forsøg viser det sig, at den bedste model er at lade den latente faktor følge en ARIMA(1,0,0)×ARIMA(1,0,0) model

$$(1 - \varphi_1 B^{12})(1 - \Phi_1 B^{12})\,\lambda_t = \varepsilon_t$$

Antal smittede blandt homoseksuelle mænd ligger på et højere niveau og er i de senere år steget langt mere end de andre tidsrækker. Derfor adderes en lokal trend til de model, der er den samme for de to andre rækker.

Sæt X_1 til kvadratroden af antal smittede heteroseksuelle mænd, X_2 kvinder og X_3 homoseksuelle mænd.

$$X_{1t} = \lambda_t + w_{1t}$$

$$X_{2t} = \lambda_t + w_{2t}$$

$$X_{3t} = \lambda_t + \beta_t + w_{3t}$$

Modellen for trendbidraget, β_t er

$$\beta_t = \beta_{t-1} + \xi_t,$$

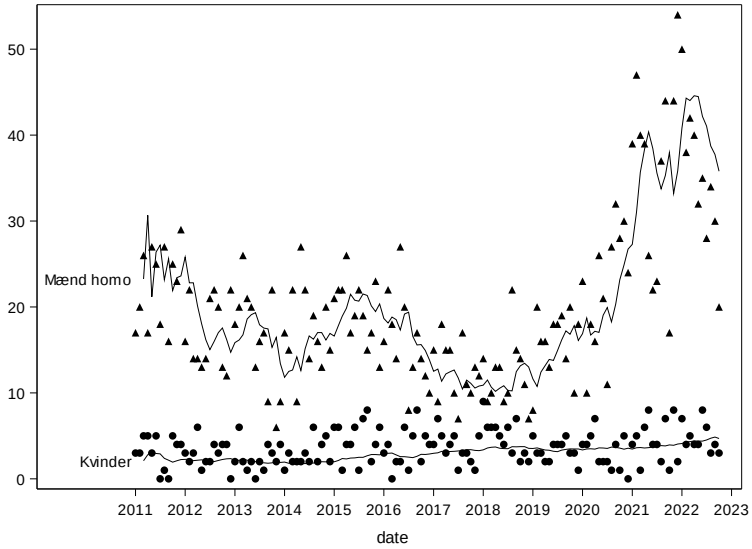
hvor ξ_t er en hvid støj.

I denne model estimeres parametrene til

$\varphi_1 = 0.99$, $\Phi_1 = 0.99$, $var(\xi_t) = 0.00029$

De to autoregressive parametre er tæt på en, men en model med differenser gik galt.

Figur 7 Forudsigelse af antal syfilissmittede for kvinder hhv. mænd smittet af mænd



Figur 7 viser modellens tilpasning til antal smittede kvinder og homoseksuelle mænd, efter at der er transformeret tilbage. Det ses, at niveauet for homoseksuelle mænd er steget meget siden 2019, men det beskrives udmærket af den lokale trend, selvom variansen på ξ_t faktisk ikke er signifikant forskellig fra nul.

Analyse af de tre tidsrækker for gonorre

Observationerne for både homo- og heteroseksuelle mænd samt for kvinder følger alle en markant opadgående trend. For heteroseksuelle mænd og for kvinder estimeres hældningerne til samme værdi. Det viser sig imidlertid, at den opadgående trend for homoseksuelle mænd ikke er lige så konstant lineær, som for de to øvrige rækker, men snarere kommer i 'ryk'. Derfor beskrives den med en lokal trend.

Den bedste model er at lade den latente faktor følger samme model som for syfilis

$(1 - \varphi_1 B)(1 - \Phi_1 B^{12})\, \lambda_t = \varepsilon_t$

Antal smittede blandt homoseksuelle mænd ligger på et højere niveau og er i de senere år steget langt mere end de andre tidsrækker. Derfor beskrives denne tidsrække ved at tilføje en lokal trend til modellen for de to andre rækker.

Modellen bliver for X_1, heteroseksuelle mænd, X_2, kvinder og X_3, homoseksuelle mænd.

$X_{1t} = \lambda_t + t + w_{1t}$

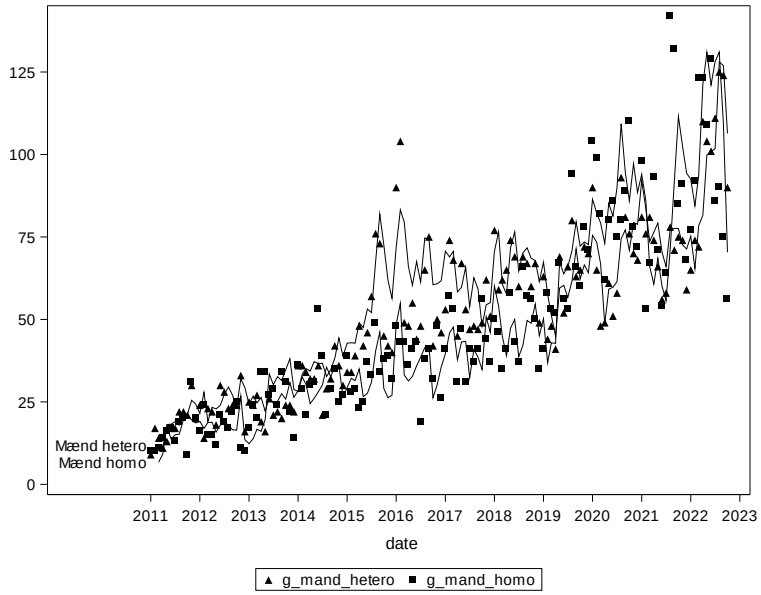
$X_{2t} = \lambda_t + t + w_{2t}$

$X_{3t} = \lambda_t + \beta_t + w_{3t}$

hvor der regresseres på tidsindekset t for heteroseksuelle mænd og for kvinder, mens β_t betegner trendbidraget for homoseksuelle mænd. Parametrene estimeres til

$\varphi_1 = 0.93$, $\Phi_1 = 0.66$, var(ξ_t) = 0.00033 og hældningen n er $\beta = 0.00034$

Figur 8 Forudsigelser af antal smittede med gonorre for hhv. mænd smittet af mænd og mænd smittet af kvinder



Figur 8 viser modellens tilpasning til antal smittede hetero- og homoseksuelle mænd, efter at der er transformeret tilbage til antal smittede. Det ses, at niveauerne stiger for begge rækker i nogenlunde samme takt; men i 2016 - 2017 og igen efter 2020 er niveauet for antal smittede homoseksuelle mænd tydeligt forhøjet. Det beskrives udmærket af den lokale trend, selvom variansen på ξ_t faktisk ikke er signifikant forskellig fra nul.

En samlet model for de tre kønssygdomme

De foregående afsnit viste, at alle rækkerne, otte i alt, beskrives ved sæson ARIMA modeller med forskellige trendled tilføjet. I det følgende afprøves en ide til at samle de tre modeller for hver sin kønssygdom til en enkelt model for alle de otte rækker under et. Som en grundmodel anvendes ARIMA modellen, der blev anvendt for gonorre tidsrækkerne, da denne model er uden differensdannelser, og den indeholder kun autoregressive og sæsonautoregressive led, der ikke i sig selv tager hensyn til trends. Den suppleres derfor med trendled svarende til tidsrækkernes udvikling.

Modellen for de enkelte tidsrækker bliver med de estimerede parametre som vist nedenfor.

Syfilis, kvinder:
Denne tidsrække indeholder ikke en trend, så den modelleres blot med sæson ARIMA modellen, der går igen for alle de otte tidsrækker. Modellen er med de estimerede parametre indsat

$(1 - 0.92B)(1 - 0.91B^{12})\ \lambda_t = \varepsilon_t$

Syfilis, mænd smittet af en kvinde:
Denne model er helt identisk med modellen for syfilis, kvinder.

Syfilis, mænd smittet af en mand:
Sæson ARIMA modellen bruges igen, men den suppleres med en lokal trend, der giver en synlig forbedring af fittet, med varians estimeret til 0.00031, der dog ikke er signifikant.

Gonorre, kvinder:
Sæson ARIMA modellen bruges igen, men den suppleres med en signifikant, $p<0.0001$, opadgående trend med hældning $\beta = 0.0015$ på dagsbasis.

Gonorre, mænd smittet af en kvinde:
Sæson ARIMA modellen bruges igen, men den suppleres med en signifikant, $p<0.0001$, opadgående trend med hældning $\beta = 0.0011$ på dagsbasis.

Gonorre, mænd smittet af en mand:
Sæson ARIMA modellen bruges igen, men den suppleres med en lokal trend, der giver en synlig forbedring af fittet, med estimeret varians 0.00020, der dog ikke er signifikant.

Klamydia, kvinder
Sæson ARIMA modellen bruges igen, men den suppleres med en signifikant, $p<0.0001$, opadgående trend med hældning $\beta = 0.0014$ på dagsbasis.

Klamydia, mænd
Sæson ARIMA modellen bruges igen, men den suppleres med en signifikant, $p<0.0001$, opadgående trend med hældning $\beta = 0.0018$ på dagsbasis.

Fittet af den samlede model for de tre kønssygdomme

Der er ikke plads til at demonstrere modellernes fit ved hjælp af plottede grafer i denne fremstilling, men Figur 9 viser et andet aspekt af fittet. Figur 9 viser autokorrelationsmatricen for de otte residuale tidsrækker for lag nul, et og tolv.

Figur 9 Residualautokorrelationsmatricen for de otte tidsrækker

Variable/Lag	*0*	*1*	*12*
RESIDUAL_kq_kvinder	+++++...	.......-	++......
RESIDUAL_kq_mand	+++++-..	.......-	++......
RESIDUAL_gq_kvinder	++++.--.	..++-.-.	....
RESIDUAL_gq_mand_hetero	++++.-..	..++.-..	....
RESIDUAL_gq_mand_homo	++..+...	....+...	....-...
RESIDUAL_sq_kvinder	.---.+..	---...+.	....
RESIDUAL_sq_mand_hetero	..-...+.	.....+..	....
RESIDUAL_sq_mand_homo	.......+	..-.....	....

** is > 2*std error, - is < -2*std error, . is between*

Figur 9 viser, at der er signifikant restautokorrelation ved lag 1 for en del par af residualrækker. Enkelte af disse autokorrelationer er numerisk i nærheden af 0.4, men langt de fleste har p-værdier kun lige under signifikansgrænsen på 5%. Ved lag 12 er der høje autokorrelationer, lidt over 0.5, for antal klamydiatilfælde for de to køn, hvilket betyder, at sæsonstrukturen for klamydia er signifikant forskellig fra sæsonstrukturen i de andre sygdomme.

Vælgerundersøgelser – styrker og svagheder

Peter Linde, peter@brede.dk

Baggrund

Jævnligt kan man læse følgende formulering i aviserne: *Ifølge en repræsentativ vælgerundersøgelse lavet af xxx (fx Gallup, Voxmeter, Epinion, Megafon, …) for yyy (fx DR, TV2, Altinget, Ritzau, Politiken, Berlingske, …) siger danskerne….* Ofte fulgt op med et forbehold om en såkaldt tilfældig statistisk fejl på 2,5%. Resultaterne står generelt til troede i den offentlige debat og de systematiske forskelle, der er mellem valgundersøgelser, tilskrives ofte tilfældige fejl. Tal er tyranni – gør alt numerisk og det betyder nærmest per definition det tallene beskriver er korrekt og faktabaseret.

Vælgerundersøgelser er en meget lille del af analysebureauernes omsætning, men en ret stor del af deres omtale. Så det er vigtigt, at troværdigheden for vælgerundersøgelser er stor. Mange har en interesse heri. Selvfølgelig de private firmaer, der sælger undersøgelser. Medierne, der citerer dem og bruger dem journalistisk. Eksperter og valgforskere, der spørges om deres vurdering. Og så selvfølgelig partierne og interesseorganisationer, der har deres egne undersøgelser. Og så alle os andre, der også har en mening.

Nogle gange kan man se vælgerundersøgelse beskrevet som en særlig form for survey. Uanset om dette er helt rigtigt, har de nogle forskelle fra andre undersøgelser. De samles ofte hurtigt ind med stort bortfald (og skævhed), men spørgsmålet er enkelt: *Hvilket parti vil du stemme på, hvis der var valg i morgen?* Over 1.000 svar er nærmes definitionen på, at man kan hæfte lid til resultatet. Kvantitet slår ligesom over i kvalitet, hvis antallet, der har svaret, bare er på fire (!) cifre. Nogle går måske så langt, at det udlægges som undersøgelsen dermed er repræsentativ. Metodisk har en repræsentativ stikprøve intet med antal svar at gøre. Antal svar har kun betydningen for sikkerheden. Repræsentativitet betyder, man har et mini-billede af befolkningen på **alle** parametre. Dette kan kun sikres, hvis man udvælger personerne tilfældigt. De vil så ligge så tæt på den samlede vælgerbefolkning, som man kan. Alle skal kunne vælges og med samme (kendte) udvalgssandsynlighed. Hvis man har et udsnit med vælgere, der er skævt (ikke repræsentativt for hele befolkningen), fx boligejere, bliver fejlen (biasen) ikke mindre - og repræsentativiteten bedre - af at udvælge dobbelt så mange fra det skæve udsnit. Det bliver kun mere sikkert skrævt.

Da vælgerundersøgelser publiceres hele tiden, kan man sammenligne deres resultater. Der er forskelle, der er vedvarende mellem undersøgelserne, når man sammenligner over tid. Når der er valg, er vælgerundersøgelser til eksamen. Her kan man se og spejle toppen af isbjerget, men måske ikke hvad der er nedenunder.

Mange vælgerundersøgelser spørger også til, hvad man stemte til sidste valg. Fx i sommeren 2022, hvad man stemte til folketingsvalget i 2019. Indsamlingsfirmaet har således både en viden om, hvad de udvalgte svarer de ville stemme, hvis der var valg 'i morgen', og hvad de stemte sidst, hvis de altså havde valgret og stemte. Hvis de kan huske det. Men med dette sidste forbehold opfyldt har man en reference. For man kan se, hvor godt de samlede svar genskaber det sidste valg. Det er ikke en information, der offentliggøres. Både fordi der vil være pæne afvigelser, men også fordi det er formidlingsmæssigt svært at beskrive to valg samtidigt, når det er de nye tal der interesserer. Viden om hvad vælgerne i dag svarer om sidste valg, kan man bruge til at vægte stikprøven, så den passer med sidste valg. Hvis et parti har en for høj andel, når der spørges til sidste valg, vægtes disse svar ned (tæller mindre end 1). Og hvis et parti har en for lille andel, vægtes deres svar op (tæller mere end 1). Disse ratio vægte kan genbruges, når man estimerer, hvad resultatet ville være, hvis der var 'valg i morgen'. Man har således kontrolleret for historisk kendt viden, og da der er en pæn positiv korrelation mellem de historiske svar og de aktuelle svar, reduceres skævheden (og faktisk også den tilfældige fejl). Dette kan man normalt kun gøre i undersøgelser, der har adgang til registeroplysninger om alle, der deltager i undersøgelsen **og** for hele populationen. Det skal være de samme oplysninger, så hvis den udvalgte, ikke kan huske hvad han stemte, er der en metodisk udfordring.

I vælgerundersøgelser har vi alle trin i processen i spil og udfordringer med dem alle: *Udvælgelsen, indsamlingen, bortfaldet og estimationen.*

Forbehold

Denne korte artikel har ikke som ærinde at kritisere nogle analysefirmaer frem for andre eller udpege en "vinder" og "taber" i 2022. Nedenfor vil nogle af forholdene ovenfor blive uddybet. Man kan selvfølgelig gå ind på de forskellige hjemmesider og læse dokumentationen for, hvordan man har gjort og hvad man har fortalt. Men livet er for kort til at finde fejl hos andre. Det er er også for let, for det er meget sværere at finde egne fejl og lære af dem. Men til den mundtlige fremlæggelse vil der komme lidt tabeller for at understøtte pointerne. Det kan af gode grunde ikke blive helt anonymt.

Udvælgelsen

Nogle vælgerundersøgelser baseres på et internet-panel af vælgere, der ofte spørges gang efter gang. Der udvælges fra internet-panelet, der kan være på rigtig mange tusinder personer. Panelerne suppleres løbende. Man kan ofte endda tilmelde sig dem, og der er normalt præmier, så man fastholder deltagelse i panelet. Rekrutteringen kan også ske gennem andre undersøgelser. Bedst selvfølgelig fra undersøgelser med høj kvalitet. Nøglen til kvaliteten af panelet er rekrutteringen. Fx er en medlemsundersøgelse i Coop nok bedre, end en undersøgelse af bilejere. Uanset hvordan man rekrutterer vil der dog være en skævhed allerede i udvælgelsen. Det er kun et spørgsmål om, hvor stor den er.

Man kan også udvælge fra databaser over telefonnumre. Der er databaser med alle telefonnumre, og man kan også genere numre tilfældigt og prøve om der forbindelse. I "gamle dage" var det mere let. Hver familie havde én telefon og 15% havde ikke - lidt forenklet. Hvis man udspurgte én person i hver valgt husstand, skulle man kun vægte for antal personer i husstanden (hvis der boede to i husstanden havde man den halve sandsynlighed for at blive valgt i forhold til en ét personers husstand). Og bagefter prøve at korrigere så godt som muligt for de 15% uden telefon. Nu om dage er verden anderledes. Hver person har måske nok en telefon, så det skal ikke på samme måde vægtes for antal personer og telefoner i husstanden. Samtidig har mange mere end én telefon. En arbejdstelefon, en privat telefon og måske en med taletidskort. De vil bliver overrepræsenteret. Fx har unge flere telefoner end ældre, arbejdsaktive flere end dem uden arbejde, rige flere end fattige osv.

Man kunne også udvælge fra CPR registeret, og fx bagefter finde telefonnummeret eller en e-mail. Offentlige institutioner kan bruge e-Boks, hvis det er en undersøgelse om deres område. Nogle tror, at private analysefirmaer ikke har adgang til CPR stikprøver, men det har de i CPR loven. De skal opfylde en række rimelige sikkerhedskrav, respekterer hvis man ikke vil deltage, og fortælle hvorfra de har oplysningen. CPR registeret er nøglen til at udvælge rigtige repræsentative stikprøver, og reelt den eneste. Alle analyse-enheder (offentlige som private) kan få adgang, så der er nok kun en udfordring. Det koster mere i investering, dataindsamlingen og drift end fx internet-paneler. Hvis man har en CPR baseret stikprøve, kan man få den vægtet til hele befolkningen efter bortfald hos Danmarks Statistik, sådan at den opregnet svarer til fx indkomstfordelingen, uddannelse, stilling, etnisk baggrund, civilstand, boligform og antal børn. Forhold der normalt betyder meget mere og er mere skævt fordelt i bortfaldet end køn, alder og geografi, som man selvfølgelig også kan korrigere for. De, der kun gør det sidste, er omvendt ikke i nærheden af noget der reelt kvalitetsforbedrer langt de fleste undersøgelser.

Indsamlingen

Hvordan man indsamler betyder også noget. Det skulle gerne svare så meget til valgsituationen som muligt. Her får man en stemmeseddel med partierne skrevet på. Ikke alle kan mellem valg huske, hvilke partier der er opstillet, og navnet på nye partier behøver ikke være alle-mands-viden. Internet besvarelse er mere fortrolig end når man udspørges af en person. Hvis internetbesvarelsen skyldes man er udvalgt fra et internet-panel, har man stadig denne skævhed. Men hvis det er over e-Boks, er der det muligt at sikre en repræsentativ udvælgelse.

Bortfaldet

Alle undersøgelser har skævhed fra bortfaldet. Det er generelt de mest integrerede, aktive og velfungerende i bred forstand, etniske danskere, dem med højst uddannelse

og højst indtægt, der svarer mest. Det betyder, at vælgerundersøgelser med kort indsamlingsperiode og stort bortfald, udfordres ved det er grupperne ovenfor, der er overrepræsenteret, og at estimatet af valgresultatet afspejler det. Det er en række partier, der ofte bliver underrepræsenteret pga. af bortfaldet. Socialdemokratiet er et af dem. Nogle mener, at de har et ekstra gear i slutspurten, men en helt naturlig forklaring kunne være metodiske udfordringer i bortfaldet. Nogle gange undervurderes socialdemokratiet ikke – fx når det går tilbage i de vælgergrupper, der underrepræsenteres i meningsundersøgelser.

Som nævnt tegner vælgerundersøgelser et for rosenrødt billede af virkeligheden. Indtægtsniveauet vil fx blive overvurderet med omkring 5%. Hvis der er diskussion af en undersøgelse, der siger, der er 4% misbrug eller 10% der har dårligt helbred, kan man normalt være sikker på, det er større. Man kan prøve at opregne herfor, men det kan kun løse en del af problemet. Derfor skal man selvfølgelig opregne alligevel – enhver kvalitetsforbedring er et plus.

Estimationen

Den sidste store kilde til bias er estimationen. Ikke hvis man gør det rigtigt, men hvis man gør det forkert eller datagrundlaget ikke kan bruges. To eksempler herpå.

I vælgerundersøgelser kan man opregne med et ratio-estimat for sidste valg. Det giver meget god mening, hvis man svarer det, man rent faktisk stemte på. Lige efter valget i 2019 er svaret om stemmen ved sidste valg mere korrekt, end når der er gået 4 år. Der har været andre valg imellem, og man har måske for to år siden skiftet parti og er ikke helt sikker på, om man allerede sidste gang stemte på ens nye parti. Det sker let med et parti vælgerne i stigende omfang støtter op om. Så hvis fx parti X ved sidste valg i 2019 fik 10% af stemmerne, og er i fremgang og i virkeligheden i 2022 står til 15%, kan man opleve, at fx 12% mener de (allerede) i 2019 stemte på X. Der er også andre korrektioner, men denne ville alene betyde at man i 2022 estimerer partiet til: 15%*10/12=12,5%. Så ratio-estimatet kan, hvis forudsætningen ikke holder, risikere at undervurdere partier i fremgang og overvurdere partier i tilbagegang, hvis man husker forkert om, hvad man stemte sidst. Vælgerundersøgelser har en tendens til at blive dårligere, jo længere tid der er gået siden sidste valg.

Det andet eksempel er opregning for fx uddannelse. Hvis opregningen bygger på, hvad der står i uddannelsesregisteret for hele befolkningen og hvad der i samme register er registeret om dem, der har svaret i den konkrete undersøgelse, er en sådan standardisering effektiv til at tage en del af det "rosenrøde" af undersøgelsen. Men hvis opregningen er baseret registeroplysninger for hele befolkningen og egen oplysning om uddannelse, går det galt. Ca. 10% oplyser, at de har en højere uddannelse end den der står i uddannelsesregisteret, og ca. 5% en lavere uddannelse. Det kan godt være rigtigt i den forstand af efteruddannelse (også på jobbet) i fx IT ikke står i registeret, ligesom

man kan have en uddannelse som geograf fra ens ungdom, man ikke har brugt i mange årtier, fordi man nu fungere som automekaniker. Men det går galt i opregningen. Egen oplysningen om ens uddannelse betyder generelt, at der bliver for mange i undersøgelsen med en høj uddannelse i forhold til det faktiske i registeret. Alle de med høj egen oplysning om uddannelse vil derfor blive vægtet ned. Også dem der har en høj uddannelse. Faktisk mere end deres øgede deltagelse i undersøgelsen (mindre bortfald) trækker estimatet af økonomiske forhold op. Som nævnt vil en undersøgelse med et normalt bortfald kunne overestimere økonomiske forhold, fx indkomst, med 5%. Men hvis man vægter med egen oplysning af uddannelse mod befolkningens registeruddannelse, er der eksempler på det modsatte – konkret at man ender med samlet at undervurderer indkomsten med 5%.

Konklusion

Churchill er gammel statistiker og skulle engang have sagt: Man kan kun handle på baggrund af statistik man selv har manipuleret, eller ved hvordan er indsamlet og beregnet. Det stiller alle, der samler data ind, overfor en opgave med at dokumentere. På hjemmesiderne kan man se, om det er lykkedes. Ellers træffer vi forkerte konklusioner.

Ved valget i 2022 var en af overraskelserne for alle meningsundersøgelser, at socialdemokraterne gik frem og fik et bedre valg end i 2019. Faktisk det bedste i 20 år. Selvfølgelig kan man give æren til socialdemokraternes særlige ever i slutspurten. Det syntes socialdemokraterne måske er positivt. Og andre en belejlig forklaring. Men det kunne måske også skyldes noget metodisk i, hvordan der udvælges, indsamles og estimeres.

Ingen undersøgelse er bedre end sit svageste led.

Nogle kilder fra egne artikler, så ansvaret alene er forfatterens

Symposiet 2020: *Repræsentativitet i paneler med løbende udskiftning. En løsning, der virker universelt, når populationen ikke er statisk.*
Symposiet 2019: *Analyse af vægtede stikprøver*.
Symposiet 2018: *Bias når surveys opregnes med egen oplyst uddannelse mod registre – hvor galt kan det gå?*
Symposiet 2016*: Hvad betyder den ophørte forskerbeskyttelse ved sammenligninger over tid.*
Symposiet 2015: *Repræsentativitet i web-paneler. Hvor skævt og kan det delvist løses?*
Symposiet 2014: *Non-response adjustment by registers and estimation of variance by replicate weights*
Symposiet 2014: *Hvad er repræsentativitet og hvad betyder det at non-response de sidste årtier har været stigende?*

Learn SAS®
Boost Your Career

More Career Opportunities

SAS is a skill employers want to see on résumés. In the past year, more than 218,000 job listings included SAS as a desired skill. (Emsi Burning Glass, October 2021)

Higher Salaries

Tech Republic named SAS as one of seven data science certifications to boost your résumé and salary.

Ongoing Skills Development

Technology is constantly evolving, and so is SAS. We'll help you keep your skills sharp and relevant with new courses and supportive communities.

FREE FOR STUDENTS
Use your university mail to sign up

SAS® Skill Builder for Students

With SAS® Skill Builder for Students, you can develop analytical skills that will boost your career opportunities and help you land your future job. Log in to SAS Skill Builder for Students to access free software and online courses, pursue valuable certifications, prep for exams & much more.

Sign up today - And boost your career with SAS!

www.sas.com/skillbuilder

All in one place - Available 24/7.

Access free software, Take Free Courses & Kickstart Your Career!