Report
1/2010

Revelation of Tax Evasion by Random Audits Report on Main Project, Part 2

Anders Berset
Erling Eide
Harald Goldstein
Paul Gunnar Larssen
Jack-Willy Olsen

Stiftelsen Frischsenteret for samfunnsøkonomisk forskning
Ragnar Frisch Centre for Economic Research

Report 1/2010

Revelation of Tax Evasion by Random Audits
Report on Main Project, Part 2

Anders Berset
Erling Eide
Harald Goldstein
Paul Gunnar Larssen
Jack-Willy Olsen

Abstract: Firms in three sectors have been subject to random audits by auditors of the Norwegian Tax Administration. The auditing has been carried out according to a detailed procedure securing that all auditors do all controls and file the results in the same manner. The auditing has been carried out in two steps, a simple and cheap control at step 1 and a comprehensive control at step 2. A test shows that the information obtained by the simple controls provides some indication of tax evasion revealed at step 2. Logistic regression analyses have been employed to test hypotheses about the effects on tax evasion of various characteristics of firms (size, age, location, use of external auditors etc.)

Keywords: Tax evasion, Random audits

Contact: www.frisch.uio.no

Report from the project "Revelation of tax evasion by random audits" (2142), funded by the Norwegian Research Council and the Norwegian Tax Administration.

ISBN 978-82-7988-092-9
ISSN 1501-9721

Contents

Summary

The project proposal *Revelation of tax evasion by random audits* was planned to consist of three parts: a Preparation Study, a Pilot Project, and a Main Project.[1] The first two parts and Part 1 of the Main Project have been carried out earlier. Several of the goals of the overall project have been reached in these parts, in particular the development of an audit strategy and a coherent system of registration. In addition, our approach has been tested on a limited number of firms.[2]

The main purpose of the present Part 2 of the Main Project has been to estimate tax evasion within some selected sectors and to investigate to which extent evasion in these sectors is dependent on various characteristics of firms.

A main feature of the audit strategy has been to carry out audits in two steps.[3] The first step consists of a not very time consuming, formal audit. At a second step, the firms for which the formal audit *indicates occurrence of tax evasion* are subject to a more comprehensive tax audit ("bokettersyn"). One purpose of this procedure is to investigate to which extent the (cheap) formal audits may reveal tax evasion.

A main goal has been to determine which firm characteristics that have a significant effect on the probability of disclosures of tax evasion and on the amounts disclosed. These factors are then included as explanatory variables in more parsimonious prediction models that may be used to estimate the expected probability of evasion and the expected amounts evaded.

The sectors selected are *joinery installation, retail sale of hardware, paint and glass, photographic activities, desiners activities*. The number of firms audited is 467 at step 1 and 191 at step 2. At step 2, tax evasion of various types (mistakes in reported net income, incorrect use of rates of value added tax, or unrecorded sales) were disclosed. Tax evasion was disclosed in 32 firms. Audits were restricted to activities related to sales.

The characteristics of tax evading firms are found to differ between the various types of evasion. There is, however, a certain tendency that firms in the most centrally located municipalities evade more than firms in other municipalities.

*

The present summary of Part 2 of the Main Study is presented in English, whereas the detailed elements of the study are found in the Appendix.

The main elements of the study are presented in Section 1 below. Section 2 describes how data were obtained. Section 3 develops a model relating various firm characteristics to a variable that *indicates* tax evasion at step 1. Section 4 estimates models that relate firm characteristics to the *probability* that tax evasion is revealed at step 2. Section 5 contains

[1] Application of 20.1.2003 to the Norwegian Research Council.

[2] See the reports: Revelation of tax evasion by random audits – Report on the Preparation Study, The Ragnar Frisch Centre, 26. juni 2005, and Revelation of Tax Evasion by Random Audits, Report on Main Project, Part 1, Ragnar Frisch Centre for Economic research, Report 1/2009.

[3] This strategy was developed at Oslo fylkesskattekontor in the pilot project and further tested in Part 1 of the Main Project, see "Revelation of Tax Evasion by Random Audits – Report on Main Project Part 1", Report 1/2009, Ragnar Frisch Centre for Economic Research.

estimates of tax evasion. The main conclusions are listed in Section 6. Section 7 informs about the staff involved in the project and its costs. In Section 8, we present our own evaluation of the project execution.

*

Data on individual tax payers has been made anonymous by "Skattedirektoratet" before the statistical analyses have been carried out.

One should note that the auditing carried out in this project is very different from the procedures ordinarily used within the Tax Administration. Our results are thus different from what traditional auditing would give.

The present report does not evaluate to which extent the Tax Administration will employ the results of the project in their ongoing activities.

1 Introduction

The previous parts of the overall project have demonstrated that reliable data cannot be obtained without using a substantial amount of auditing resources. In order to save on such resources it was decided to rely on the audit strategy (described in section 2 below) developed in the Pilot study of the overall project. A detailed procedure of how audits should be carried is formalised in a PC-program. The idea has been that all the auditors, when auditing, should be obliged to follow the same procedure and register their findings in boxes supplied by the PC-program.

The previous parts of the overall project suggested that available resources were sufficient only to audit firms in a few sectors. It was decided in Part 2 of the project to study four sectors: *joinery installation, retail sale of hardware, paint and glass, photographic activities, and designers activities.*

These sectors include both activities where evasion previously has been revealed, and activities that so far have been investigated only to a modest degree by the Tax Administration.

In order further to limit the requirement of audit resources only activities related to sales have been audited.

A main feature of the audit strategy has been to carry out audits in two steps.[4] The first step consists of a not very time consuming, formal audit. At a second step, the firms for which the formal audit *indicates occurrence of tax evasion* are subject to a more comprehensive tax audit ("bokettersyn"). A main goal of this two-step procedure has been to investigate to which extent the (cheap) formal audits may reveal tax evasion.

A main purpose has been to determine which explanatory factors that have a significant effect on the probability of disclosures of tax evasion and on the amounts disclosed. These factors are then included as explanatory variables in more parsimonious prediction models that are used to estimate the expected probability of evasion and the expected amounts.

Like the study in Part 1, the present study is mainly of an exploratory kind. In statistics, it is common to distinguish between exploratory and confirmatory studies. The exploratory element in our study consists mainly in our search for models of prediction (including rather few explanatory variables) that can be used to explain data. These models are not chosen a priori, but obtained by a more or less systematic search in the data at hand. The reason for this approach is that the number of potential explanatory variables is rather large compared to the amount of data that reasonably can be obtained. The number of possible prediction models or explanatory models (with significant explanatory variables) is high and data does not contain enough information to distinguish one from the other.

Simulations have demonstrated that when the space of potential prediction models is large, the probability of spurious significances is great. On the other hand, experience indicates that if there is a relationship between the response variable and some explanatory variables, there is a rather great chance such a relationship will be included in some of the models that are not rejected by data. The implication is that even if an explanatory variable is strongly significant (low p-value) in any of the prediction models studied, the only conclusion to be drawn is that

[4] This strategy was developed at Oslo fylkesskattekontor in the pilot project and further tested in Part 1 of the Main Project, see "Revelation of Tax Evasion by Random Audits – Report on Main Project Part 1", Report 1/2009, Ragnar Frisch Centre for Economic Research.

there is *some* evidence for the variable to be of importance, not that there is *strong* evidence. In order to conclude that the evidence is strong, the relationship has to be tested against new data (in another study).

1.1 Variables and statistical analyses

1.2 Variables

1.2.1 Response variables

Step 1
At step 1, the response variable *Z* is a variable that represents weaknesses in the firms' internal controls and in their quality of accounts. It is a hypothesis of our study that such weaknesses are related to tax evasion. For convenience, we use the term *hint of evasion* to characterize *Z*, without implying that evasion has in fact been revealed. The value of this (dual) variable is determined through the following procedure. Based on detailed reports of the auditors, we have computed a summary statistic, a *"technical" evaluation* ("MaksAvPoeng"), indicating the quality of internal routines and books. In addition, the auditors have carried out an *overall evaluation* of whether they expect a firm to evade tax. The auditors have ranked the firms according to a scale from 1 to 4, where 1 indicates satisfactory routines and books and 4 indicates very serious mistakes/faults. Their *overall evaluation* based on these "marks" ("samlet vurdering", *SV*). The value of Z is then

Z = 1 if MaksAvPoeng $\geq$ 0.2 or SV = 3 or SV = 4
Z = 0 if not

Z is employed as a screening variable to sort out those firms for which tax evasion is most likely to be revealed at step 2. At step 2 all firms for which Z=1 were audited, as well as a random selection of those for which Z=0.

Step 2
At step 2, we distinguish between *disclosure of tax evasion Y* (a dual variable) and the corresponding *disclosed amount X*. If there is a *disclosure of tax evasion* at step 2, the amount of evasion is positive. The *disclosed amounts* are measured by the differences between correct amounts as determined by the auditors and the amounts reported by the firms.

We distinguish between three types of X (X_1, X_2, X_3) and corresponding*Ys*: (Y_1, Y_2, Y_3).

X_1 is the *amount of change in net income* except mistakes consisting of wrong periods of registration and mistaken use of value added rates.
X_2 is the *amount of change value added tax* (VAT) caused by mistakes in the use of VAT rates etc.
X_3 is the *amount of unregistered sales* (sales not included in books).

1.2.2 Explanatory variables (exogenous covariates)

The explanatory variables are of several types, all but one dummies. (The dummies are equal to zero if the "ifs" are not satisfied.)

Sector

Snekker	Dummy = 1 if *joinery installation*
Jernv	Dummy = 1 if *retail sale of hardware, paint and glass*
Fotograf	Dummy = 1 if *photographic activities*
Design	Dummy = 1 if *desiners activities*

Region

Ost	Dummy = 1 if Tax Region East,
Sor	Dummy = 1 if Tax Region South
Vest	Dummy = 1 if Tax Region West
Midt	Dummy = 1 if Tax Region Central Norway
Nord	Dummy = 1 if Tax Region North

Type of firm

AS	Dummy = 1 if corporation
ENK	Dummy = 1 if sole proprietorship

Number of employees

A0	Dummy = 1 if zero employees
A1	Dummy = 1 if 1-3 employees

Age

Nyreg	Dummy = 1 if firm has existed in less than 4 years
	Dummy = 0 if firm has existed at least 4 years

External accountant

R	Dummy = 1 if external accountant

Type of municipality where the firm is located

Komsentral	Dummy = 1 if the municipality is among the most centrally located
Komtjenest	Dummy = 1 if service sectors dominate in the municipality
KSminKTJ	= Komsentral – Komtjenst

Sales

Oms_3	Dummy = 1 if sales < 300 000 NOK
Oms3_10	Dummy = 1 if 300 000 NOK < sales < 1 000 000 NOK
Oms0_10	Dummy = 1 if sales < 1 000 000 NOK

A main purpose has been to determine which explanatory factors that have a significant effect on disclosures, and use these factors – and only these – to construct parsimonious models that in the future can be used to *predict* disclosures.

2 Data

The data, which were collected in 2008, are with one exception related to the firms' activities in 2007, see Table 1. At step 1, 467 firms were subject to the formal audit, whereas 191 were subject to the more comprehensive audit at step 2. The number of firms for which evasion was indicated at step 1 ($Z = 1$) is given in the last column.

Table 1 No. of firms audited in step 1 and step 2 with percentage of hints of evasion

	No. of observations		
Sector	Step 1	Step 2	Hint of evasion at step 1 (%)
45.42 – Joinary installation	224	99	22.8 (51/224)
52.46 – Retail sale of hardware, colour and glass	97	42	9.3 (9/97)
74.81 – Photographic activities	66	25	22.7 (15/66)
74.87 – Designers activities	80	25	21.3 (17/80)
Sum	**467**	**191**	**19.7 (92/467)**

2.1 Comprehensive audit, step 2

Among the191 firms have been audited at step 2, the auditors have proposed changes in net income (X_1), in VAT (X_2), or in sales (X_3) for 32 of them, see column 1 of Table 2 for details. Table 2 also shows average, median, minimum and maximum values of these variables. For X_2, there in one negative observation (-1 775 NOK). This observation is excluded in the statistical tests, and the number of observations are reduced from 14 to 13. Furthermore, one extreme value (611 774 NOK) has a substantial effect on the average etc. for X_2. In some calculations below, the extreme value is excluded when the effect of X_2 is studied, see the last but one row of Table 2.

Table 2 Various types and amounts of tax evasion disclosed at step 2

	No of firms	Average	Standard error	Median	Min. value	Max. value
X_1	24	114 456	161 709	67 431	8 898	685 787
X_2	14	57 768	160 070	9 463	-1 775	611 774
$X_2 > 0$	13	62 349	165 649	10 000	477	611 774
X_2 without extreme value	12	16 563	14 304	9 463	477	39 787
X_3	16	57 225	70 648	23 503	678	217 600

2.2 Other explanatory variables

Data required for the remaining variables listed in section 1.2 are obtained from various files available in Skattedirektoratet.

2.3 Data file and descriptive statistics

Together with data from existing files in Skattedirektoratet, the data obtained from the audits has been included in a comprehensive file made available for analysis. All data on this file has been made anonymous. Descriptive statistics based on the data file is given in the Appendix

3 The probability distribution of hint of tax evasion (*Z*) at step 1

The first task has been to study the probability distribution of Z, see 3.1. The main purpose of establishing the probability distribution of Z, is to control for skewness in the distributions of disclosures (*Y*) and disclosed amounts (*X*) caused by the use of *Z* as a screening variable at step 1.

The research strategy has been to establish first a full model for *the probability of hint of evasion*, i.e. $P(Z=1)$, including all the exogenous variables listed above. From this full model, an explorative search has been carried out in order to establish a more parsimonious prediction model that can be used in the following statistical tests.[5] A number of sub-models have been studied, using various methods of excluding and including covariates. Details of the preferred prediction model are included in Table 3. An LR-test against the full model indicates that almost nothing is lost by excluding all the other covariates, see Table 3.

It turned out that the combined variable KSminKJT = Komsentral –Komtjenst produces a better fit than Komsentral and Komtjenst separately. The regression coefficient of KSminKJT is negative, which means that the *probability of hint of evasion* is (i) lowest among municipalities that are among the most centrally located and that are not dominated by service industries (KSminKJT=-1), and (ii) highest among municipalities that are dominated by service industries ant not centrally located (KSminKJT=1).

One notes that the probability of hint of evasion is higher in Tax Region East than in other tax regions, and that firms with external accountant have lower probability of hint of evasion than other firms.

[5] The search procedure is described in "Revelation of Tax Evasion by Random Audits – Report on Main Project Part 1", Report 1/2009, Ragnar Frisch Centre for Economic Research. In the literature a number of criteria have been proposed in order to choose among the various sub-models, such as the p-values of estimated coefficients, likelihood-ratio (LR) testing, and various information criteria. Among possible information criteria, the common AIC, Akaike's information criteria, and his Bayesian modification, BIC, have been used.

Table 3 Regression results (logistic regression) for hint of evasion at step 1 (Z)

Explanatory variables	Full model		Prediction model	
	Coefficient	p-value	Coefficient	p-value
AS	----	----	----	----
ENK	0.7008	0.178	----	----
Ost	1.4555	0.002	1.1491	0.000
Sor	0.3840	0.386	----	----
Vest	0.3194	0.494	----	----
Midt	-0.2823	0.534	----	----
Nord	----	----	----	----
Snekker	0.2173	0.584	----	----
Jernv	0.0722	0.905	----	----
Fotograf	0.2905	0.530	----	----
Design	----	----	----	----
Nyreg	-0.2170	0.605	----	----
A0	0.8445	0.241	----	----
A1	0.4182	0.527	----	----
Komtjenest	0.6605	0.029	----	----
Komsentral	-0.6270	0.071	----	----
KSminKJT	----	----	-0.5675	0.010
R	-0.6915	0.017	-0.4922	0.066
Oms0_3	0.3020	0.491	----	----
Oms3_10	0.3746	0.351	----	----
Constant	-2.9098	0.000	-2.3471	0.000
No. of observations	456		467	456
Log-likelihood	-196.3419		-206.1368	-198.8258
-2 log LR				4.9677
p-value reduced vs. full model				0.986

The zeros in the four rows in the full model are restrictions in order to avoid multicollinearity.

4 The effect of various covariates of on the probability of disclosures at step 2

Sections 4.1-4.3 present the probabilities of disclosure of the three types of evasion (Y_1, Y_2, and Y_3) given the screening result at step 1.

4.1 Probability of disclosures of tax evasion, given Z

4.1.1 Probability of disclosure of the type *change in net income* (Y_1), given *Z*

At step 2, of the 191 firms audited, tax evasion of the type *change in net income* was disclosed in 24 firms (12.6%). Logistic regressions, similar to those describe above, were carried out for a full model and a number of sub-models. The full model is described in the Appendix. The explorative research led to two prediction models, the results of which are given in Table 4. Prediction model 1 includes only two explanatory variables: sole proprietorship (*ENK*) and Tax Region East (*Ost*). Prediction model 2 includes sole proprietorship (*ENK*) and the most centrally located municipalities (*Komsentral*). The second column of Table 2 shows (for comparison) some of the results for the full model.

Table 4. Regression results (logistic regression) for disclosure of the tax evasion type "change in net income" (Y_1).

Explanatory variables	Full model		Prediction model 1		Prediction model 2	
	Coefficient	p-value	Coefficient	p-value	Coefficient	p-value
ENK	0.3577	0.772	1.5706	0.040	1.6025	0.036
Ost	1.3686	0.172	1.0942	0.021		
Komsentral	0.7295	0.282			1.0003	0.039
Constant	-3.2210	0.178	-3.5134	0.000	-3.8115	0.000
No. of observations	188		191	188	191	188
Log-likelihood	-59.5926		-65.5267	-64.6523	-65.7465	-65.1101
-2 log LR				10.1194		11.0350
p-value reduced vs. full model				0.860		0.807

The two prediction models are not very different as far as log-likelihood testing is concerned. When excluded variables are included one at a time, the p-values of the included variables are about 0.20 or greater. Consequently, none of the excluded variables seems to add to the explanation of Y_1.

One reason for the two prediction models to be rather similar probably is that almost all the municipalities (35 out of 37) in Tax Region East are among those most centrally located.

An interesting (and rather somewhat unexpected?) result is that the variable *Z* (*hint of tax evasion* at step 1), which is included in the full model, does not seem to have any effect on Y_1. If, in particular, we add Z to prediction model 1, and to prediction model 2, the coefficient of Z obtains p-values of 0.934 and 0.488, respectively. This result strengthens the conclusion that the estimated probabilities of disclosure at step 2 are the same whether or not a hint of tax evasion is obtained at step 1.

Table 5 shows some results for prediction model 1. Column 5 contains the calculated probability of (the auditor's proposal of) *change in net income* at step 2. Because Z according to the prediction model has no effect on the probability of disclosure at step 2, the probabilities in the table are the same regardless of the value of Z, and the probabilities in the table may be interpreted as prevalences.

Table 5 Probability of disclosure of type change in net income, given hint of evasion at step 1 for prediction model 1 – all 4 sectors

Hint of evasion at step 1	Type of firm	Tax Region East	Relative frequency	Probability	Lower unilateral 95% confidence limit
Hint of evasion, Z=1	ENK	Yes	0.27 (7/26)	0.300	0.187
		No	0.13 (6/45)	0.125	0.080
	Not ENK	Yes	0.00 (0/1)	0.082	0.023
		No	0.09 (1/11)	0.029	0.009
Not hint of evasion, Z=0	ENK	Yes	0.33 (2/6)	0.300	0.187
		No	0.13 (7/54)	0.125	0.080
	Not ENK	Yes	0.25 (1/4)	0.082	0.023
		No	0.00 (0/44)	0.029	0.009

In the Appendix a table similar to Table 5 is presenter for prediction model 2, and the same conclusion is drawn about Z.

4.1.2 Probability of disclosure of the type *VAT mistakes* (Y_2), given Z

At step 2, of the 191 firms audited, tax evasion of the type *VAT mistakes* was disclosed in 14 firms (7.3%), see Table 2 for details. Logistic regressions, similar to those describe above, were carried out for a full model and a number of sub-models. The explorative search led to a prediction model including only *Midt, Z and Komsentral* as explanatory variables. In this case, at variance with the result for *change in net income*, the screening variable *Z* turned out to be highly significant, see Table 6.

Table 6 Probability of disclosure of type VAT mistakes, given hint of evasion at step 1 for prediction model 1 – all 4 sectors

Hint of evasion at step 1	Region	Municipality most central	Relative frequency	Probability	Lower unilateral 95% confidence limit
Hint of evasion, Z=1	Central	Yes	0.50 (2/4)	0.40	0.17
		No	0.00 (0/6)	0.16	0.06
	Not central	Yes	0.18 (7/38)	0.17	0.10
		No	0.06 (2/35)	0.06	0.02
Not hint of evasion, Z=0	Central	Yes	0.00 (0/8)	0.09	0.03
		No	0.09 (2/22)	0.03	0.01
	Not central	Yes	0.03 (1/39	0.03	0.01
		No	0.00 (0/39)	0.01	0.00

4.1.3 Probability of disclosure of the type *unregistered sales*, given *Z*

At step 2, of the 191 firms audited, tax evasion of the type *unregistered sales* (Y_3) was disclosed in 16 firms (8.4%). Logistic regressions, similar to those describe above, were carried out for a full model and a number of sub-models. The explorative research led to two prediction models. Prediction model 1 includes as explanatory variables *Z, external accountant* (*R*)*, and most centrally located municipalities* (*Komsentral*), and prediction model 2 includes *Z, sales between 0 and 1 mill NOK* (*Oms0_10*), *and Komsentral* as explanatory variables. (Here, the variable *sales between 0 and 1 mill NOK* is interpreted as a measure of the size of the firms, i.e. small firms.)

The Appendix includes tables similar to Table 6 for both prediction models. Prediction model 2 fits slightly better than prediction model 1. Some results for prediction model 2 are given in Table 7. Unregistered sales seem to be most common in firms with sales less than 1000 000 NOK, in particular in firms that do not have an external accountant.

Note that the probabilities in Table 7 are the same whether or not there was an hint of evasion at step 1. This means that the probabilities may be interpreted also as prevalence probabilities.

Table 7 Probabilities of disclosure of type unregistered sales, given hint of evasion at step 1 (Z=1) for prediction model 2 – all 4 sectors

Hint of evasion	External accountant	Centrally located municipality	Sales under 1 mill. NOK	Relative frequency	Probability	Lower one-sided confidence limit
Hint of evasion	Yes	Yes	Yes	0.10 (2/21)	0.12	0.06
			No	0.17 (1/6)	0.02	0.00
		No	Yes	0.00 (0/18)	0.04	0.01
			No	0.00 (0/8)	0.01	0.00
	No	Yes	Yes	0.43 (6/14)	0.31	0.18
			No	--- (0/0)	0.06	0.01
		No	Yes	0.17 (2/12)	0.12	0.05
			No	0.00 (0/2)	0.02	0.00
No hint of evasion	Yes	Yes	Yes	0.10 (2/21)	0.12	0.06
			No	0.17 (1/6)	0.02	0.00
		No	Yes	0.00 (0/18)	0.04	0.01
			No	0.00 (0/8)	0.01	0.00
	No	Yes	Yes	0.43 (6/14)	0.31	0.18
			No	--- (0/0)	0.06	0.01
		No	Yes	0.17 (2/12)	0.12	0.05
			No	0.00 (0/2)	0.02	0.00
Total				0.09 (16/188)		

4.2 Prevalence probabilities of disclosures of tax evasion

4.2.1 Probability of disclosure of type *change in income* (Y_1), controlled for Z

As explained above, there was no evidence in data that the screening at step 1 had any effect on the probability of *change in net income.* Consequently, the estimated prevalence probabilities are equal to the probabilities in Table 5.

The probability of disclosure of *change in net income* seems to be considerably higher in Tax Region East than in other regions. The same holds true for the most centrally located municipalities compared with other municipalities, as well as for firms with sole proprietorship compared to other types of firms.

4.2.2 Probability of disclosure of type *VAT mistakes* (Y_2), controlled for Z

In this case, the screening at step 1 had a significant effect. Combining results described in above the vector of explanatory variables that seems to have an effect on the probability of disclosure of VAT mistakes is

U = (Ost, Midt, R, ENK, Komsentral, Komtjenst)

The probability, controlled for Z =1 is

$$P(Y_3 = 1|U) = q \cdot p_1 + (1-q) \cdot p_0$$

where
q = $P(Z=1|$ *Ost, R, ENK, KS*smin*KTS*$)$
$p_0 = P(Y_2 = 1|$ *R, Komsentral,* $Z = 0)$
$p_1 = P(Y_2 = 1|$ *R, Komsentral,* $Z = 1)$

The following results were obtained: The estimated probabilities are highest for Tax Region Central Norway, somewhat lower in Tax Region East and lowest in the remaining tax regions. The probabilities are slightly higher in firms without external accountant than those with external accountant. The combination of characteristics that have the highest probabilities in all regions are firms with without external accountant and belonging to the most centrally located municipalities.

4.2.3 Probability of disclosure of type *unregistered sales* (Y_3), controlled for *Z*

Two possible prediction models were obtained. In prediction model 1, Z is an explanatory variable, and consequently one has to control for *hint of evasion* at step 1. In prediction model 2, Z is not included, and such control is not required.

Prevalence probabilities for prediction model 1
A procedure similar to the one described in section 4.2.2 gave the following results: The probabilities for Tax Region South is about 25 % higher than those for the other tax regions. The highest probabilities are obtained for the firms from the most centrally located municipalities and without external accountant.

Prevalence probabilities for prediction model 2
In this prediction model Z is not included, and consequently the probabilities in Table 7 may be interpreted as prevalence probabilities.

In this case, it turns out that the variable *sales under 1 mill. NOK*, which indicates the size of the firms, has a clear effect on the probability; the probability is 6-7 times as high as for other firms. Moreover, this variable seems to eliminate the effect of *hint of evasion* at step 1 (which makes it unnecessary to control for the screening at step 1). In addition, this variable seems to eliminate the effects of Region, net income and Komtjenst, explanatory variables included in prediction model 1.

A comparison of the two models using the information criteria AIC and BIC further strengthen prediction model 2 against prediction model 1.

5 Estimation of expected *amounts* of changes, given changes (*Y*)

Subsections 5.1-5.3 present the estimated expected amount of changes (X) in net income, VAT, and sales, respectively, given that changes have been revealed (Y=1)

5.1 Estimation of expected amount of *change in net income* X_1

At step 2, of the 191 firms audited there are 24 observation of X_1. An initial analysis is based on generalised linear modelling (GLM) with a Gamma distribution and log link function.[6]

Two extreme observations have caused some estimation problems. It has been found that these observations have a substantial influence of the effect of *Z* on the amount of change in net income. A comprehensive investigation of this relationship concludes that there is scant, if any, evidence for *Z* to have an effect on the amount of change in net income. A number of possible prediction models without *Z* have been studied. Some of them have been rejected because of over fitting, a well-known problem in ordinary regression analyses when the number of explanatory variables approaches the number of observations.

In one of the preferred prediction models, only sole proprietorship (*ENK*) and external accountant (*R*) are included as explanatory variables. For this model, the expected amounts of changes in net income are given in Table 8. For ease of reference the various combinations of ENK and External accountant are numbered, see the first column of the table. (Similar numbering is applied in tables 9 and 10 below.)

Table 8 Expected amount of change in net income (X_1), given change in net income. Confidence limits based on robust standard error. Bootstrap (BCa) confidence limits in brackets. No. of observations in brackets in column 4.

Category	ENK	External accountant	Observed average amount of change in net income (1000 NOK)	Expected amount of change in net income (1000 NOK)	95 % confidence interval	
					Lower limit	Higher limit
1	Yes	Yes	154 (14)	155	81 (81)	295 (309)
2		No	68 (8)	68	36 (36)	127 (133)
3	No	Yes	33 (1)	30	20 (--)	45 (--)
4		No	12 (1)	13	8 (--)	21 (43)
5	Category 3 and 4 together		22 (2)	22	12 (--)	43 (--)

One may observe that the expected amounts of changes are almost equal to the observed ones. The reason is that almost all observations are from the group of firms of sole proprietorship (*ENK*). Because of few observations in groups 3 and 4, together with the comparatively many observations in group 1 and 2, the estimates in group 3 and 4 cannot be considered as reliable. The same holds true for the interval of confidence for these groups.

There seems to be some evidence for the amounts of chance in net income, given change, to be somewhat higher for firms having external accountant than for those which do not have external accountant.

[6] The modelling is described in "Revelation of Tax Evasion by Random Audits – Report on Main Project Part 1", Report 1/2009, Ragnar Frisch Centre for Economic Research.

In another prediction model where *zero employees* are substituted for *external accountant*, rather similar results are obtained.

It is of interest to note that 21 of the 24 observations are related to firms characterised by sole proprietorship (*ENK*) and zero employees (*A0*).

5.2 Estimation of expected amount of *change in VAT* (X_2)

There are only 13 observations of X_2, too little for comprehensive regression analyses. Nevertheless, two prediction models representing the expected amount of changes in VAT have been developed. As expected, the results illustrate the problem of too few observations. In particular, the problem of over fitting is predominant.

Some results of the preferred prediction model is given in Table 9. The esimates included in brackets are particularly unreliable (because of no observations in the corresponding group).

Table 9 Expected amount of change in net income (X_2), given change in net income. Confidence limits based on robust standard error. Bootstrap (BCa) confidence limits in brackets. No of observations in brackets in column 5.

Category	ENK	Newly registered	Sales below 1 mill. NOK	Observed average amount of change in net income (1000 NOK)	Expected amount of change in net income (1000 NOK)	95 % confidence interval	
						Lower limit	Higher limit
1	Yes	Yes	Yes	39 (2)	39	37	40
2			No	--- (0)	(153)	(73)	(325)
3		No	Yes	5 (6)	5	3	9
4			No	21 (3)	21	12	35
5	No	Yes	Yes	-- (0)	(594)	(144)	(2447)
6			No	-- (0)	(2368)	(572)	(9802)
7		No	Yes	-- (0)	(80)	(18)	(365)
8			No	320 (2)	320	86	1192

5.3 Estimation of expected *amount of unregistered sales* (X_3)

There are 16 observations of unregistered sales. All these firms, but one, belong to a group characterised by having sole proprietorship (*ENK*) and zero employees (*A0*). A number of possible prediction models for this group of firms (called "B group") have been investigated by the usual explorative procedures. In a preferred model only Tax Region East (*Ost*) and sector *hardware etc*. (*Jernv*) is included as explanatory variables.

No evidence was found for Z to have any influence on the amount of unregistered sales.

The expected amounts of unregistered sales are given in Table 10.

Table 10 Expected amount of unregistered sales (X_3), given change in net income. Confidence limits based on robust standard error. Bootstrap (BCa) confidence limits in brackets (4000 replications). No. of observations in brackets in column 4.

Category	Tax Region East	Sector hardware	Observed average amount of change in net income (1000 NOK)	Estimated amount of change in registered sales (1000 NOK)	95 % confidence interval	
					Lower limit	Higher limit
1	Yes	Yes	-- (0)	2	1 (--)	5 (--)
2		No	21 (5)	21	15 (--)	30 (--)
3	No	Yes	7 (2)	7	3 (--)	18 (--)
4		No	88 (9)	88	49 (43)	160 (145)
5	Category 2 and 4 together		64 (14)	64	36 (33)	116 (111)
6	All categories (1-4)		57 (16)	57	31 (32)	105 (103)

The three largest observations (about 200 000 NOK) are all in category 4, whereas the remaining observations are well below 100 000 NOK. Because of the low number of observations, it is reasonable to assume that the estimated amounts of changes in registered sales are overestimated in category 4 and underestimated in the other categories. Taking all categories together (implying an assumption that the expected amounts are equal in categories 1-4) the expected amount of change is 57 000 NOK. Lacking a good theory of how often extreme values appear, no better prediction seems possible to find for group B.

5.4 Estimation of expected *amount of change in net income* disregarding cases of unregistered sales (X_4)

There are 14 observations of changes in net income that are not related to unregistered sales. Also for this group, an explorative search for prediction models has been carried out. Like in several subsections above, there is no evidence that Z has any effect on the amount of change in this net income of the defined type. Scant evidence indicates that the expected changes are higher for newly registered firms located in Tax Region East than for other firms.

6 Some conclusions

- The purpose of screening in step 1 has been to investigate whether a rather simple audit may indicate that tax evasion will be revealed in the more comprehensive audit at step 2. No evidence was found of the screening effect on the tax evasion measures of *changes in net income*[7] and of *unrecorded sales*. No evidence was found either with regard to the *probability of disclosure* or with regard to the *amount* of change in net income given

[7] Excluded in this measure are changes in net income caused by the use of wrong periods in bookkeeping and by use of wrong percentage of value added tax.

change. As for the change in the value added tax, however, there is evidence for the screening to have an effect on the probability of disclosure, but not on the *amount* evaded given disclosure of evasion.

- The screening seems to have had less effect for the sectors studied in the present Part 2 of the Main study compared with the sectors studied in Part 1. One reason might of course be that there are in fact differences between the two groups of sectors. Another explanation might be that a somewhat more lax screening in Part 2 resulted in a larger proportion of firms in step 2 that do not evade tax.
- A remarkable find is that almost all cases of disclosures (*change*) occurred among firms having sole proprietorship, without income and without employees (called *group A*) .
- In the sectors *joinery installation, photographic activities, and desiners activities* 70-80% of the firms belonged to Group A. Only 9% of the firms in sector *hardware* belonged to this group.
- The predominance of firms in group A, small sets of data, and some extreme observations of disclosed amount of evasion produces a great degree of over-fitting in models predicting *amounts of change* in income etc., given *change*. In such cases the estimates of expected changes in income etc. were equal to the observed averages among the firms in the group. Consequently, information *between* groups became rather meagre, and the prediction of expected changes in income etc. in groups with only a few observations had little sense.
- The data on the size of the changes in *income etc.* indicate that most observations are rather moderate whereas there are a few extreme values. This tendency was clear in Part 1 of the Main Study and is also present in Part 2. Such a distribution will in our case, in which the estimated *amounts of changes in income etc*. tend to be equal to the observed averages, produce underestimation in subgroups that do not contain extreme observations and overestimation in subgroups with few, but one or several extreme observations. Consequently, it is not recommendable to aggregate estimates of changes in *income etc.* except for group A. However, in a situation where information *between* groups prevails, so that the prediction is based on a larger number of observations, aggregation within groups might be acceptable.
- A consequence of the tendency of overestimation or underestimation in subgroups seems to be that the most reliable estimation of expected amounts of changes in income etc. is obtained by employing data of all subgroups together, the results of which are given in Table 11.

Table 11 Estimated expected amount of change in net income, VAT and sales, given change, for all groups. Bootstrap standard error and confidence limits (4000 replications). No of observations in brackets.

			95% confidence interval	
Type of change	Estimated expected amount of change, given change (1000 NOK)	Standard error (1000 NOK)	Low limit of confidence	High limit of confidence
Change in net income	114 (24)	32	69	213
Unrecorded sales	57 (16)	17	31	99
Change in net income except because of unrecorded sales	131 (14)	51	55	283
Change in value added tax because of wrong VAT rate	62 (13)	44	14	245

- The tendency that sporadic extreme observations appear in data should be considered in future modelling of the size of changes in *income etc.*, given change. The class of Gamma distributions employed both in Part 1 and Part 2 will to some degree take this tendency into account, but not to a satisfactory extent for out type of data.1
- The possibility of over-fitting does not pose the same problem for the prediction of the probability of change as for the prediction of the amount of change in income etc, given change. One reason is that the number of observations (191 in step 2) is much higher. Moreover, exact prediction is not relevant in a logistic regression because such a prediction would imply that some regression coeffitions would be plus or minus infinity, cases that must be discarded before carrying out the regression analysis.
- With some reservations, there is evidence for the probability of disclosure of *change in income etc.* to be higher in Tax Region East than in other regions. This probability of disclosure is higher also in the more centrally located municipalities compared with other municipalities. Furthermore, the probability of disclosure is higher for firms with sole proprietorship than for other firms. The highest probability found (.23) is estimated for firms with sole proprietorship in the more centrally located municipalities. The lowest probabilities are found for firms other than with sole proprietorship in less centrally located municipalities.
- As for disclosure of *change in value added tax related to mistakes of VAT rates etc. related to sales* (without increase in net income) the highest estimated probabilities are found for Tax Region Central Norway, somewhat lower for Tax Region East, and lowest for Tax Region South, Tax Region West, and Tax Region North. Firms without external accountant have a somewhat higher probability (20%) than those with external accountant. The combination of firm characteristics that have the highest probability in all regions is firms without external accountant and located in the more centrally located municipalities.

- The probability of *disclosure of unrecorded sales* is found to be highest among firms in the most centrally located municipalities, without external accountant and sales less than 1 million NOK. The lowest estimated probabilities are found for firms outside the most centrally located municipalities and with sales above 1 million NOK.
- The probability of disclosure of change in net income of other types than unrecorded sales appears to be highest among firms of sole proprietorship in Tax Region East with sales above 1 million NOK (probability 0.44 with lower confidence limit 0.20). The lowest estimates of these probabilities were found for firms without sole proprietorship not in Tax Region East and with sales less than 1 million NOK. It is worth noticing that all firms in this category, except one, were not newly registered (older than 4 years), and most of them (11 out of 14) had external accountant.

7 Staff and costs

7.1 Staff

The project has been carried out by Erling Eide, University of Oslo and the Frisch Centre, ass. professor Harald Goldstein, University of Oslo, Paul Gunnar Larssen, Jack-Willy Olsen and Anders Berset at the Tax Administration. A number of auditors at the Tax Administration have participated in the production of data.

7.2 Costs and resources employed

The project has been financed partly by the Norwegian Research Council and partly by the Tax Administration (Skattedirektoratet). The Norwegian Research Council has covered the participation by researchers at the Ragnar Frisch Centre for Economic Research and University of Oslo (500 000 NOK), whereas the Tax Administration ("Skattedirektoratet") has covered auditing and monitoring of the audit procedure.

The time used by the Tax Authority is as follows:

Step 1: Two man-days (MD) per audit, 467 audits	934 MD
Step 2: Four MD per audit, 191 audits	766 "
Development and evaluation of audit strategy, training of auditors, monitoring of audits.	70 "
	1760 MD

Here, the time used at step 1 is an estimate based on previous experience, whereas the time used at step 2 is an average obtained from actual time registration. (The time used to develop the audit system in previous parts of the RA project is estimated to 200 man-days.)

Assuming the costs per MD to be 2000 NOK, the Tax Authority's total cost amounts to 3 520 000 NOK.

8 Summary of project execution

The strategy of auditing, a system of registration of audit results, the establishment of data files, model building and tests has been developed and carried out according to the project plan. Some

concern might be raised by the fact that some audits have been carried out by rather unexperienced auditors.

The statistical work has been rather demanding. Because of a rather limited data set and few observed hints of evasion at step 1 of the audit procedure, it has been difficult to obtain statistically significant effects on evasion of various characteristics of firms.

Prediction models explaining tax evasion as functions of certain characteristics of firms have been developed and tested. Some estimates of the magnitude of tax evasion have been obtained. Because of a rather limited number of observations, the estimates are not very precise. We believe, however, that the method we have developed will produce more precise estimates when more data becomes available.

The audit strategy we have developed seems to be suitable for various types of industries, and the (somewhat imprecise) estimates indicate the variation in tax evasion among industries.

We have decided not to try to estimate the effects of sanctions. The data required seem to be out of reach.

Harald Goldstein
Revidert februar 2010

Random Audit Project

Statistisk analyse av data fra 2007 - innhentet 2008

Analyse av endringer av typen

- **"nettoinntekt bortsett fra feilperiodiseringer og feil bruk av mva-satser",**
- **"endring av merverdiavgift relatert til avgiftsfeil på salgsområdet (uten økning i nettoinntekt)" og**
- **"påvist uteholdt omsetning (kontrollmelding o.l. på salg som ikke er bokført)"**

0. Innhold

1. Innledning

Denne rapporten er en oppfølging av analysen rapportert i Frisch rapport 2009/1, "Revelation of Tax Evasion by Random Audits. Report on Main Project. Part 1" (kalt **FR** nedenfor) av data innhentet i 2007, og er basert på nye data innhentet i 2008, samt nye bransjer. Dataene innhentet i 2007 stammer i hovedsak fra 2006 (og kalles "2006-dataene"), mens dataene innhentet i 2008 stammer i hovedsak fra 2006 og 2007 (og kalles "2007-dataene" nedenfor).

Det metodiske grunnlaget for den statistiske analysen for 2007-dataene bygger stort sett på appendiks F (kalt **HR** nedenfor) og appendiks G (kalt **SR** nedenfor), begge i **FR**.

Tabell 1.1 Oversikt over bransjer og antall observasjoner for data innhentet i 2008 og 2007

		Antall observasjoner		
Data innhentet	Bransje	Trinn1	Trinn2	Funn trinn 1 %
2008	45.42 - Snekkerarbeid	224	99	22.8 (51/224)
	52.46 - Butikkhandel med jernvarer, fargevarer og glass	97	42	9.3 (9/97)
	74.81 - Fotografvirksomhet	66	25	22.7 (15/66)
	74.87 - Designvirksomhet	80	25	21.3 (17/80)
	Sum	467	191	19.7 (92/467)
2007	51.4 - Engroshandel med klær, sports- og fritidsutstyr mv.	74	18	9.9 (7/71)
	60.240 - Godstransport på vei	120	34	4.5 (5/112)
	74.700 - Rengjøring	97	31	21.3 (20/94)
	Sum	291	83	11.6 (32/277)

Merk at dataene innhentet i 2007 mangler noen (14) observasjoner for screenings-indikatoren (Funn på trinn 1).

2007-dataene har generelt samme struktur som 2006-dataene. Det totale utvalget er gitt ved det såkalte trinn-1-utvalget. Virksomhetene i trinn-1-utvalget undersøkes ved en relativt rask og rimelig screening-test, mens alle virksomhetene på trinn 2 utsettes for full materiell kontroll. Trinn-2-dataene er et utvalg fra trinn-1-enetene delvis basert på utfallet av screening-testen beskrevet ved en indikator for "*funn på trinn 1*".

Screening-testen består først og fremst av en undersøkelse av formale sider ved virksomheten som kan hentes ved en relativt rask intervjuundersøkelse. På bakgrunn av denne beregnes en såkalt *MAV*-skåre mellom 0 og 1 som et uttrykk for risikoen for at en materiell kontroll skal avdekke grunnlag for endring av nettoinntekt eller mva. I tillegg gir revisor en mer subjektivt basert skåre, "*samlet vurdering* (811)", fra 1 til 4. Detaljer om disse skårene kan leses i Frisch rapport 2009/1, "Revelation of Tax Evasion by Random Audits. Report on Main Project. Part 1" (**FR**).

Screening-testen for 2007-dataene er en utvidelse av testen for 2006-dataene. For 2006-dataene ble bare MAV-skåren benyttet og "funn på trinn 1" definert som $MAV \geq 0.3$. Analysen av 2006-dataene viste at screeningstesten hadde en klar effekt i de tre bransjene som ble valgt samt at den subjektive skåren ville ha hatt en effekt. I håp om å fange opp flere "risiko-virksomheter" ble funn-1-kriteriet utvidet i to retninger for 2007-dataene. Dels ble *MAV*-kriteriet senket til 0.2, og et kriterium basert på den subjektive skåre (skåre minst lik 3) ble lagt til. Denne utvidelsen er antakelig hovedgrunnen til at funn-1-prosentene i tabell 1.1 er noe høyere i 2007-dataene enn i 2006-dataene. En av konklusjonene i denne rapporten er at screeningen synes å ha hatt mindre effekt for de nye bransjene utvalgt for 2007-data enn for

de tre bransjene i 2006-dataene. Dette kan naturligvis skyldes forskjeller mellom de to forskjellige (disjunkte) bransje-settene, men det foreligger også en mulighet at utvidelsen av screeningskriteriet har vært for liberal. Siden de to bransje-settene ikke har noen bransjer felles, inneholder dataene dessverre ikke informasjon til å kunne teste denne muligheten.

Utvalgsplanen med screening på trinn 1 skaper skjevheter i utvalget på trinn 2 som må kontrolleres for. Denne kontrollen er integrert i metodikken utviklet i HR og begrunnet der. Begrunnelsen vil derfor ikke bli gjentatt i denne rapporten.

I tillegg til bransje er materialet trukket stratifisert over fem regioner

- Skatt nord (Finnmark, Troms og Nordland)
- Skatt Midt-Norge (Nord-Trøndelag, Sør-Trøndelag, Møre og Romsdal)
- Skatt vest (Sogn og Fjordane, Hordaland, Rogaland)
- Skatt sør (Vest-Agder, Aust-Agder, Telemark, Vestfold og Buskerud)
- Skatt øst (Oslo, Akershus, Østfold, Hedmark og Oppland)

Stratifiseringen har i praksis vært noe mer detaljert enn som bestemt av bransje (tabell 1.1) og region. For eksempel utvalget fra bransje 52.46 er videre trukket stratifisert etter undergruppene

52.461	Butikkhandel med bredt utvalg av jernvarer, fargevarer og andre byggevarer
52.462	Butikkhandel med jernvarer
52.463	Butikkhandel med fargevarer
52.464	Butikkhandel med trelast
52.469	Butikkhandel med byggevarer ikke nevnt annet sted

Forbehold 1. På grunn av det relativt begrensete materialet vil vi ignorere slike substrata som i jernvarehandel-bransjen, og anta at stratifiseringen er definert ved bransje (som i tabell 1.1) og region. Dette innbærer homogenitetsantakelser over ignorerte substrata. Ved den modellbaserte tilnærmingen (i motsetning til designbasert), som er valgt her (jfr. HR avsnitt 8), betyr dette at vi antar at observasjonsvektorene er uavhengige og identisk fordelte innenfor hvert av de 20 strataene definert ved region og bransje fra tabell 1.1. Spesielt bygger framstillingen på antakelsen at utvalget innenfor hvert bransjestratum (inklusive substrata) er trukket rent tilfeldig innenfor hver region.

Forbehold 2. Denne studien, i likhet med HR og SR, bærer sterkt preg av å være av såkalt eksplorativ type. I statistikk skiller man gjerne mellom *eksplorative* og *bekreftende* (confirmatory) studier. Det eksplorative elementet hos oss er først og fremst det at prediksjonsmodellene (med relativt få forklaringsvariable), som vi bruker som grunnlag for tolkning av data, ikke er kjent eller valgt på forhånd (a priori), men valgt basert på en mer eller mindre systematisk leting i foreliggende data. Med et relativt stort antall av potensielle forklaringsvariable som vi har her, betyr dette at det foreligger et stort antall av mulige prediksjons- eller forklaringsmodeller (med signifikante forklaringsvariable) som data ikke har informasjon nok til å kunne diskriminere imellom. Mange av disse signifikansene kan være spuriøse (ikke reelle - dvs. kun tilstede i foreliggende data, men sannsynligvis ikke i nye data trukket fra samme populasjon). Simuleringsstudier viser at det er høy sannsynlighet for at spuriøse signifikanser oppstår når rommet av potensielle prediksjonsmodeller er stort. På den

annen side er det også erfaring for at hvis en sammenheng mellom responsen og noen forklaringsvariable er reell (i populasjonen), så er det relativt stor sjanse at en god letestrategi vil oppdage det i betydning av at relasjonen med høy sannsynlighet vil være med i klassen av kandidater for prediksjonsmodeller som ikke forkastes av data. Dette innebærer at selv om en forklaringsvariabel er sterkt signifikant (liten p-verdi) i en av prediksjonsmodellene foreslått nedenfor, så kan vi ikke si mer enn at det er en *viss* evidens i data for at variabelen er betydningsfull, men ikke grunnlag for si at det er *sterk* evidens (som den lave p-verdien nominelt skulle tilsi). For å kunne konkludere med sterk evidens trengs bekreftende studier der den aktuelle sammenhengen konfronteres med nye data.

For eksempel, hvis en av bransjene hadde vært felles for 2006- og 2007-dataene, ville vi kunne ha gjennomført en bekreftende analyse av screeningens betydning i denne rapporten, men siden det ikke finnes felles bransjer er analysen av screeningen fortsatt til en stor grad eksplorativ.

Konfidensgrenser. I tabellene for estimerte sannsynligheter i avsnitt 4 er det, istedenfor vanlige 95% konfidensintervall, oppgitt nedre 95% konfidensgrenser - siden jeg antar at en nedre konfidensgrense har større interesse enn en øvre. En ensidig nedre 95% konfidensgrense for en parameter, p, er den observerte verdien av en stokastisk variabel, A, som oppfyller $P(A \leq p) \approx 0.95$. Merk at den ensidige nedre konfidensgrensen ligger litt høyere enn den tilsvarende nedre verdien i et tosidig konfidensintervall, slik at vi vinner noe informasjon på denne måten. For eksempel, tabell 4.11 viser at sannsynligheten for endring av nettoinntekt for en tilfeldig ENK-virksomhet trukket utenfor Skatt øst, er estimert til 0.125 med ensidig nedre 95% konfidensgrense 0.080. Den tilsvarende nedre verdien i et tosidig 95% konfidensintervall er 0.074.

2. Variable

.

2.1 Responsvariable

Trinn 1:

På trinn 1 har vi bare en respons nemlig “funn på trinn 1” indikert ved

$$Z = \begin{cases} 1 & \text{hvis } MAV \geq 0.2 \text{ eller } \textit{samlet vurdering}(811) \text{ får verdi 3 eller 4} \\ 0 & \text{ellers} \end{cases}$$

der MAV (*MaxAvVerdi*) er en skåre på skala fra 0 til 1, beregnet på grunnlag av revisors vurdering på trinn 1 av en rekke formelle forhold. Z utgjør screening-variabelen som ble benyttet til å effektivisere utvalget (øke avdekking-sannsynlighetene) på trinn 2.

Trinn 2:

$$Y = \text{"endring"} = \begin{cases} 1 & \text{hvis materiell kontroll (trinn 2) fører til endring} \\ 0 & \text{ellers} \end{cases}$$

$$X = \text{“endringstall”} = \text{ størrelsen på beløpet som endres } \begin{cases} > 0 \text{ hvis } Y = 1 \\ = 0 \text{ hvis } Y = 0 \end{cases}$$

X og Y opptrer i tre versjoner. Ingen av versjonene omfatter feilperiodiseringer:

- X_1 omfatter endringer i nettoinntekt bortsett fra feilperiodiseringer og feil bruk av mva-satser. Y_1 er en tilsvarende avdekkings-indikator (= 1 hvis $X_1 > 0$ og = 0 ellers).
- X_2 omfatter endring av merverdiavgift relatert til avgiftsfeil på salgsområdet (uten økning i nettoinntekt). Y_2 er den tilsvarende avdekkings-indikatoren.
- X_3 omfatter påvist uteholdt omsetning (kontrollmelding o.l. på salg som ikke er bokført). Y_3 er den tilsvarende avdekkings-indikatoren.

I tillegg trengs en indikator for når $X_1 > X_3$,

$$Y_{1a} = \begin{cases} 1 & \text{hvis } X_1 > X_3 \\ 0 & \text{ellers} \end{cases}$$

Merk at påvist uteholdt omsetning nødvendigvis impliserer endring i nettoinntekt. Av dette følger at $Y_3 \le Y_1$ (eller, med andre ord, at $(Y_3 = 1) \Rightarrow (Y_1 = 1)$) alltid gjelder).

2.2 Forklaringsvariable (eksogene kovariater)

Bransje.

Snekker - Dummy = 1 for bransje, 45.42 - Snekkerarbeid**,** og = 0 ellers.

Jernv - Dummy = 1 for bransje, 52.46 - Butikkhandel med jernvarer, fargevarer og glass, og = 0 ellers.

Fotograf - Dummy = 1 for bransje, 74.84 - Fotografvirksomhet, og = 0 ellers.

Design - Dummy = 1 for bransje, 74.87 - Designvirksomhet, og = 0 ellers.

Region

Ost - Dummy = 1 for region Skatt Øst, og = 0 ellers.

Sor - Dummy = 1 for region Skatt Sør, og = 0 ellers.

Vest - Dummy = 1 for region Skatt Vest, og = 0 ellers.

Midt - Dummy = 1 for region Skatt Midt-Norge, og = 0 ellers.

Nord - Dummy = 1 for region Skatt Nord, og = 0 ellers.

Virksomhetstype

AS - Dummy = 1 for aksjeselskap og = 0 ellers.

ENK - Dummy = 1 for enkeltmannsforetak og = 0 ellers.

Antall ansatte

A*0* - Dummy = 1 for null antall ansatte og = 0 ellers.

A*1* - Dummy = 1 for en til tre ansatte og = 0 ellers.

Alder

$$Nyreg = \begin{cases} 1 & \text{Nyregistrert (eksistert i 3 regnskapsår eller færre} \\ 0 & \text{etablert (4 eller flere regnskapsår)} \end{cases}$$

Ekstern regnskapsfører

R - Dummy = 1 hvis virksomheten har ekstern regnskapsfører, og = 0 ellers.

Kommunetype

$$Komsentral = \begin{cases} 1 & \text{hvis kommunen mest sentral (jfr. SSB definisjon 2008: 3 = mest sentral)} \\ 0 & \text{ellers (dvs. 0 - 2 ifølge SSB definisjon)} \end{cases}$$

$$Komtjenest = \begin{cases} 1 & \text{hvis dominerende næringstruktur i kommunen er tjenesteyting (6-7 iflg SSB definisjon 1994)} \\ 0 & \text{ellers} \end{cases}$$

$$KSminKTJ = Komsentral - Komtjenest$$

Omsetning

Oms0_3 - Dummy = 1 for omsetning under 300 000, og = 0 ellers.

Oms3_10 - Dummy = 1 for omsetning mellom 300 000 og 1 mill., og = 0 ellers.

Oms0_10 = Oms0_3 + Oms3_10
- Dummy = 1 for omsetning under 1 mill., og = 0 ellers.

Merknader

Alle forklaringsvariable er således dikotome i denne analysen bortsett fra *KSminKTJ* som tar tre verdier, 1, 0 og -1.

Endringsprosenter. Blant de 191 virksomhetene trukket ut på trinn 2 for 2007-dataene var det 24 avdekkinger av type 1 ("endring av nettoinntekt"), 14 avdekkinger av type 2 ("endring av merverdiavgift") og 16 avdekkinger av type 3 ("påvist uteholdt omsetning"). Blant de 14 avdekkingene av type 2 var det 6 som også ga avdekking av type 1 og 8 som bare hadde avdekking av type 2.

Det var dermed i alt 16.7% (32 av 191) avdekkinger av type 1 eller 2 i trinn-2-utvalget. For 2006-dataene var den tilsvarende prosenten 26.6% (22 av 83).

Virksomhetstype konsentrerer seg hovedsakelig på typene AS (29%) og ENK (68%). De øvrige typene (3%) omfatter typene, ANS, DA, NUF og VIFE. Dette betyr at ENK og AS er praktisk talt komplementære i dette materialet, og jeg vil derfor kun bruke dummien for ENK som forklaringsvariabel nedenfor med den forståelsen at kategorien ikke-ENK hovedsakelig omfatter AS.

Antall ansatte har blitt erstattet av dummiene *A0 (0 ansatte) og A1 (1-3 ansatte).* Grunnen til det er at antall ansatte, som varierer mellom 0 og 40, har en sterkt skjev fordeling med 84% tre eller færre ansatte. En eventuell effekt av en slik variabel uttrykt ved enkelt regresjonskoeffisient kan lett bli misvisende hvis effekten er selv svakt ikke-lineær. En viss kompensasjon for dette oppnås ved gruppering.

Tabell 2.1 Frekvenstabell for antall ansatte

	Antall ansatte			
	0	1-3	>3	Sum
Abs. frekv.	313	78	76	467
%	67	17	16	100

Virksomhetens alder har blitt erstattet med dummien for nyregistrert (*Nyreg*) - dvs. alder høyst 3 år. I likhet med analysen i SR viste *Nyreg* seg å gi litt bedre tilpasning enn *alder* i de tilfeller der alder synes å ha betydning. Alder, som varierer mellom 1 og 38 år med median 11, er også karakterisert ved en sterkt høyreskjev fordeling.

Omsetning opptrer som to variable i databasen, *Sum Avgpl Oms (Post 2) 2006* og *Sum Avgpl Oms (Post 2) 2007 (pr 3-01-08-dvs 5 term)*. Den første variabelen har 18 manglende observasjoner ("missings"), og den andre 84 missings. Av de to omsetningsvariablene dannet jeg en kombinert omsetningsvariabel, *Omsetning*, som er lik den første der denne har verdi og lik den andre der bare den andre har verdi. Dette reduserte antall missings til 11. Den kombinerte omsetningsvariabelen varierer mellom 0 og 67 mill med gjennomsnitt 33 mill og median 624 000. Fordelingen er således sterkt skjev med 62% verdier under 1 mill og 38% verdier mellom 1 og 67 mill. Av samme grunn som antydet for antall ansatte ble derfor *Omsetning* erstattet av to dummier, *Oms0_3* og *Oms3_10*, (omsetning 0 - 300 000 og 300 000 - 1 mill hhv). I noen tilfeller, for eksempel når deres regresjonskoeffisienter var relativt like, ble de to erstattet med en enkelt dummy, *Oms0_10* (omsetning 0 - 1 mill). Dette er ekvivalent med å postulere at de to regresjonskoeffisientene er like.

Tabell 2.2 Frekvenstabell for omsetning

	Omsetning			
	0-300 000	300 000 - 1 mill	> 1 mill	Sum
Abs. frekv.	128	157	171	456
%	28	34	38	100

Databasen inneholder variabelen **Sum skattbar inntekt 2006.** Denne variabelen er ufullstendig i og med at den kun er registrert for etterskuddspliktige virksomheter, og er derfor utelatt fra analysen.

Databasen inneholder også en variabel, **Beløp Lønn(111A) 2006,** som mulig kandidat for forklaringsvariabel. Imidlertid inneholder denne 289 manglende observasjoner, og er derfor heller ikke tatt med.

Sensurering av hobbyvirksomheter. Det ble foretatt en sensurering midt i utvalgsplanen (mellom trinn 1 og 2), nemlig fjerning av såkalte "hobbyvirksomheter eller lignende" fra den delen av trinn-1-utvalget som ikke gir funn på trinn 1.

"Hobbyvirksomhet" er definert som "ingen ansatte eller omsetning under 100 000". Ca 5 av de 24 avdekkingene i alt på trinn 2 oppfyller denne definisjonen på "hobbyvirksomhet" slik at variabelen kunne ha potensial som forklaringsvariabel. Den er imidlertid ikke tatt med blant forklaringsvariablene siden sensureringen avskjærer muligheten for å kontrollere for screeningsskjevheten i eventuelle effekter av denne variabelen.

2.3 Nummerisk oversikt over variable

I tabell 2.4-2.7 nedenfor følger en oversikt over variablene i analysen.

Tabell 2.4 Frekvenstabell for dikotome variabler i analysen.

	TRINN 1			TRINN 2		
	Antall obs.	Antall =1	%	Antall obs.	Antall =1	%
Z	467	92	20	191	83	43
Y1	-----	-----	-----	191	24	13
Y2	-----	-----	-----	191	14	7
Y3	-----	-----	-----	191	16	8
Y1a	-----	-----	-----	191	14	7
AS	467	135	29	191	52	27
ENK	467	316	68	191	131	69
Andre virksomheter	467	16	3	191	8	4
Ost	467	84	18	191	37	19
Sor	467	99	21	191	36	19
Vest	467	95	20	191	39	20
Midt	467	92	20	191	40	21
Nord	467	97	21	191	39	20
Snekker	467	224	48	191	99	52
Jernv	467	97	21	191	42	22
Fotograf	467	66	14	191	25	13
Design	467	80	17	191	25	13
Nyreg	458	61	13	188	19	10
A0	467	313	67	191	135	71
A1	467	78	17	191	28	15
Komtjenest	467	208	45	191	89	47
Komsentral	467	237	51	191	89	47
R	467	347	74	191	137	72
Oms0_3	456	128	28	188	43	23
Oms3_10	456	157	34	188	74	39

Tabell 2.5 **Frekvenstabell for KSminKTJ = Komsentral – Komtjenest.**
(Ks = Komsentral, Ktj = Komtjenest)

TRINN 1	Ks=0, Ktj=1	Ks=Ktj	Ks=1, Ktj=0	Sum
Antall	71	296	100	467
%	15	63	21	100
TRINN 2	Ks=0, Ktj=1	Ks=Ktj	Ks=1, Ktj=0	Sum
Antall	37	117	37	191
%	19	61	19	100

Tabell 2.6 **Oversikt over endringsbeløp (*X*) på trinn 2.**

	Antall $\neq 0$	Gj.snitt	St.avvik	Median	Min	Maks
X_1	24	114 456	161 709	67 431	8 898	685 787
X_2	14	57 768	160 070	9 463	-1775	611 774
$X_2 > 0$	13	62 349	165 649	10 000	477	611 774
$X_2 > 0$ uten maks-verdien	12	16 563	14 304	9 463	477	39 787
X_3	16	57 225	70 648	23 503	678	217 600

Gruppe A. Det viser seg at undergruppen (som jeg har kalt *gruppe A*) av virksomheter som består av *enkeltmannsvirksomheter med ingen ansatte*, kommer til å spille en viktig rolle i analysen de avdekkete endringsbeløpenes størrelse gitt endring, i og med at de aller fleste endringstilfellene (28 av i alt 32 av type 1 eller 2) som ble avdekket, kommer fra gruppe A. Tabell 2.7 gir en oversikt over gruppe A i trinn-1-utvalget. En dummy for gruppe A kan oppfattes som en samspillsvariabel ved

$$GruppeA = ENK \times A0$$

Tabell 2.7 **Frekvenstabell for gruppe A over region og bransje i trinn-1-utvalget. Prosent. Absolutte tall i parentes.**

	Gruppe A	Andre	Sum
Totalt	62.3 (291)	37.7 (176)	100 (467)
Skatt øst	66.7 (56)	33.3 (28)	100 (84)
Skatt sør	70.7 (70)	29.3 (29)	100 (99)
Skatt vest	54.7 (52)	45.3 (43)	100 (95)
Skatt midt-Norge	58.7 (54)	41.3 (38)	100 (92)
Skatt nord	60.8 (59)	39.2 (38)	100 (97)

Snekkerarbeid	81.3 (182)	18.7 (42)	100 (224)
Jernvarehandel	9.3 (9)	90.7 (88)	100 (97)
Fotografvirksomhet	68.2 (45)	31.8 (21)	100 (66)
Designvirksomhet	68.8 (55)	31.2 (25)	100 (80)

3. Sannsynlighetsfordlingen for funn på trinn 1, *Z*

Sannsynlighetsfordelingen for *Z* brukes først og fremst til å kontrollere for skjevheter i fordelingene for avdekking (*Y*) og avdekkingsbeløp (*X*) som introduseres ved screenings-variabelen *Z*.

Tilsvarende strategi for etablering av full og redusert modell for $P(\text{Funn trinn 1}) = P(Z = 1)$, som beskrevet i HR , leder til resultatene gitt i tabell 3.1. I dette tilfellet synes den reduserte modellen (prediksjonsmodellen) å være ganske klart bestemt. Ingen klart "konkurrerende" prediksjonsmodeller viste seg.

Tabell 3.1 **Regresjonsresultater (logistisk regresjon) for funn på trinn 1 (*Z*)**

(Basert på utskrift Ut 9-10 i appendiks A4)

Avhengig *Z*	Full modell		Prediksjonsmodell	
	Koeff.	p-verdi	Koeff.	p-verdi
AS	----	----	----	----
ENK	0.7008	0.178	1.3919	0.000
Ost	1.4555	0.002	1.1442	0.000
Sor	0.3840	0.386	----	----
Vest	0.3194	0.494	----	----
Midt	-0.2823	0.534	----	----
Nord	----	----	----	----
Snekker	0.2173	0.584	----	----
Jernv	0.0722	0.905	----	----
Fotograf	0.2905	0.530	----	----
Design	----	----	----	----
Nyreg	-0.2170	0.605	----	----
A0	0.8445	0.241	----	----
A1	0.4182	0.527	----	----
Komtjenest	0.6605	0.029	----	----
Komsentral	-0.6270	0.071	----	----
KSminKTJ	----	----	-0.5675	0.010
R	-0.6915	0.017	-0.4922	0.066
Oms0_3	0.3020	0.491	----	----
Oms3_10	0.3746	0.351	----	----
konstant	-2.9098	0.000	-2.3471	0.000
Antall obs	456		467	456
Log-likelihood	-198.3359		-209.5705	-202.3736

-2 log LR				8.07546
P-verdi redusert vs full modell				0.779

Variabelen *KSminKTJ = Komsentral - Komtjenest* viser seg her å gi bedre tilpasning enn kommunevariablene Komsentral og Komtjenest hver for seg. Med negativ regresjonskoeffisient betyr det at det er minst sannsynlighet for funn 1 i kommuner som er mest sentrale og ikke hovedsakelig tjenesteytende (KSminKTJ = 1), og mest sannsynlig i kommuner som er hovedsakelig tjenesteytende og ikke mest sentral (KSminKTJ = -1). Andre kommuner har en sannsynlighet for funn 1 som ligger imellom. Merk at det å erstatte *Komsentral* og *Komtjenest* med variabelen *KSminKTJ* er ekvivalent med å pålegge restriksjonen at regresjonskoeffisienten for *Komsentral* er lik minus koeffisienten for *Komtjenest*.

De fire hullene i den fulle modellen er pålagt for å unngå eksakt eller nesten eksakt multikollinearitet.

Vi merker oss også at Skatt Øst har høyere sannsynlighet for funn 1 enn andre regioner, og virksomheter med ekstern regnskapsfører har lavere sannsynlighet enn andre virksomheter.

4. Betydning av eksogene kovariater for sannsynligheten for avdekking på trinn 2

Avsnitt 4.1 - 4.3 beskriver sannsynlighetene gitt utfallet av screeningen på trinn 1, mens avsnitt 4.4 - 4.8 beskriver tilsvarende prevalens-sannsynlighetene - dvs. sannsynlighetene kontrollert for screeningen på trinn 1.

4.1 Sannsynligheten for avdekking av typen "endret nettoinntekt" (indikator Y_1) gitt utfallet av screeningen på trinn 1.

Det var i alt 24 avdekkinger av denne typen av 191 i trinn-2-utvalget (12.6%).

Tilsvarende framgangsmåte som i avsnitt 6 i HR ga følgende tabell over regresjonsresultatene for avdekking, full modell og to alternative reduserte prediksjonsmodeller.

Tabell 4.1 **Regresjonsresultater (logistisk regresjon) for avdekking av typen "endret nettoinntekt" (indikator Y_1) .**

(Basert på utskrift Ut 1-3 i appendiks A1)

Avhengig Y_1	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
AS	-2.4220	0.245	----	----	----	----
ENK	0.3577	0.772	1.5706	0.040	1.6025	0.036
Ost	1.3686	0.172	1.0942	0.021	----	----
Sor	0.7033	0.500	----	----	----	----
Vest	0.0395	0.973	----	----	----	----

Midt	1.1500	0.227	----	----	----	----
Nord	----	----	----	----	----	----
Snekker	0.5697	0.462	----	----	----	----
Jernv	1.1574	0.315	----	----	----	----
Fotograf	0.7514	0.405	----	----	----	----
Design	----	----	----	----	----	----
Z	0.0985	0.866	----	----	----	----
Nyreg	-0.2229	0.775	----	----	----	----
R	-0.4639	0.398	----	----	----	----
Oms03	0.3959	0.600	----	----	----	----
Oms310	-0.4485	0.523	----	----	----	----
A0	-0.2056	0.906	----	----	----	----
A1	-0.2046	0.905	----	----	----	----
Komtjenest	0.1118	0.859	----	----	----	----
Komsentral	0.7259	0.282	----	----	1.0003	0.039
Konstant	-3.2210	0.178	-3.5134	0.000	-3.8115	0.000
Antall obs	188		191	188	191	188
Log-likelihood	-59.5926		-65.5267	-64.6523	-65.7465	-65.1101
-2 log LR				10.1194		11.0350
P-verdi redusert vs full modell				0.860		0.807

Estimatene i de reduserte modellene er basert på 191 observasjoner, mens LR testen er basert på de 188 observasjonene uten missings for den fulle modellen.

De to hullene i den fulle modellen skyldes at de tilhørende variablene (Nord og Design) er fjernet på grunn av multikollinearitet. De er dermed referansegrupper for region og bransje hhv.

Et påfallende trekk ved tabell 4.1 er at så få forklaringsvariable kommer ut som signifikante i de to prediksjonsmodellene. Dette kan naturligvis delvis skyldes begrenset informasjon i data, men også det faktum at kategorien "*endret nettoinntekt*" er inhomogen som responskategori bestående av forskjellige typer som "*uteholdt omsetning*" og "*andre typer*" viser seg å være viktig . Avsnitt 4.3 og 4.6 viser nemlig at både *ekstern regnskapsfører* og *omsetning* har betydning for de to delkategoriene av "*endret nettoinntekt*", mens tabell 4.1 viser at effekten av disse to forklaringsvariablene synes å forsvinne når vi slår sammen responskategoriene "*uteholdt omsetning*" og "*andre typer*".

De to prediksjonsmodellene er ganske likeverdige. De er nesten identiske med hensyn til likelihood-ratio testing mot full modell, samt informasjonskriteriene BIC og AIC (de siste ikke rapportert). Hvis vi føyer til en av de utelatte variablene til prediksjonsmodellen, vil p-verdien for den tillagte variabelen i nesten alle tilfeller holde seg over 0.20 og aldri under 0.17. Ingen av de utelatte variablene synes således å ha noe å bidra med til forklaring av Y_1 utover de inkluderte variablene.

En medvirkende årsak til at de to prediksjonsmodellene har så like egenskaper er trolig at nesten alle kommunene (35 av 37 i data) i Skatt øst er av den mest sentrale typen (*Komsentral* =1) i trinn-2-utvalget, jfr. tabell 4.3.

Tabell 4.3 Innebyrdes fordeling av *Komsentral* og Skatt *Øst* (trinn-2-utvalget)

	Komsentral		
Øst	0	1	Sum
0	100	54	154
1	2	35	37
Sum	102	89	191

Det er interessant at funn-indikatoren på trinn 1, *Z*, ikke synes å ha betydning for Y_1, noe som, etter sjekking, gjelder uansett hvilken av de øvrige variable som tas med i prediksjonsmodellen i tillegg. Hvis vi spesielt legger *Z* til forklaringsvariablene i prediksjonsmodell 1 og 2, får koeffisienten til *Z* p-verdi henholdsvis 0.934 og 0.488, som styrker konklusjonen at *Z* ikke har betydning for Y_1. Dermed blir de estimerte sannsynlighetene i tabell 4.4-5 de samme enten det var funn på trinn 1 eller ikke.

Når det gjelder tolkningen av tabell 4.4 og 4.5 over tilsvarende sannsynligheter (og lignende tabeller senere over sannsynligheter basert på foreslåtte prediksjonsmodeller), kan følgende sies:

- Forskjellene mellom de estimerte sannsynlighetene i tabellen kan anses som signifikante siden de er basert på signifikante effekter fra prediksjonsmodellen (under forbehold 2 fra avsnitt 1).
- Det at det ikke ble funnet evidens (igjen under forbehold 2) for at de øvrige forklaringsvariablene i den fulle modellen hadde betydning utover de valgte forklaringsvariablene, kan tolkes slik at de estimerte sannsynlighetene er representative for alle delgrupper bestemt av utelatte variable. For eksempel, sannsynligheten for avdekking av typen "endret nettoinntekt" for ENK-virksomheter i Skatt Øst er estimert til 0.300 (tabell 4.4) uansett om virksomheten har ekstern regnskapsfører eller ikke.

Tabell 4.4 Sannsynligheter for avdekking av typen " endret nettoinntekt" , betinget av funn trinn 1 ($Z = 1$**), for prediksjonsmodell 1 (fra tabell 4.1). Alle 4 bransjer.**

Funn trinn 1	Enhetstype	Region Skatt Øst	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Funn	ENK	Ja	0.27 (7/26)	0.300	0.187
		Nei	0.13 (6/45)	0.125	0.080
	Ikke ENK	Ja	0.00 (0/1)	0.082	0.023
		Nei	0.09 (1/11)	0.029	0.009
Ikke funn	ENK	Ja	0.33 (2/6)	0.300	0.187
		Nei	0.13 (7/54)	0.125	0.080
	Ikke ENK	Ja	0.25 (1/4)	0.082	0.023
		Nei	0.00 (0/44)	0.029	0.009
		Total	0.13 (24/191)		

Tabell 4.5 **Sannsynligheter for avdekking av typen " endret nettoinntekt" , betinget av funn trinn 1 (** $Z = 1$), **for prediksjonsmodell 2 (fra tabell 4.1). Alle 4 bransjer.**

Funn trinn 1	Enhetstype	Kommune Mest sentral	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Funn	ENK	Ja	0.31 (12/39)	0.230	0.158
		Nei	0.03 (1/32)	0.099	0.053
	Ikke ENK	Ja	0.00 (0/3)	0.057	0.017
		Nei	0.11 (1/9)	0.022	0.006
Ikke funn	ENK	Ja	0.13 (4/30)	0.230	0.158
		Nei	0.17 (5/30)	0.099	0.053
	Ikke ENK	Ja	0.06 (1/17)	0.057	0.017
		Nei	0.00 (0/31)	0.022	0.006
		Total	0.13 (24/191)		

Merk at enhetstype "Ikke ENK" består hovedsakelig av AS, (dvs 52 av 60 i trinn-2-utvalget).

4.2 Sannsynligheten for avdekking av typen "endring av merverdiavgift relatert til avgiftsfeil" (indikator Y_2) gitt utfallet av screeningen på trinn 1.

Det var i alt 14 avdekkinger av denne typen av 191 i trinn-2-utvalget (7.3%).

Tabell 4.6 **Regresjonsresultater (logistisk regresjon) for avdekking av typen "endret merverdiavgift" .**

(Basert på utskrift Ut 4-5 i appendiks A2)

Avhengig Y_2	Full modell		Prediksjonsmodell	
	Koeff.	p-verdi	Koeff.	p-verdi
AS	-1.5422	0.210	----	----
ENK	----	----	----	----
Ost	0.3011	0.724	----	----
Sor	----	----	----	----
Vest	----	----	----	----
Midt	1.9312	0.032	1.1817	0.090
Nord	----	----	----	----
Snekker	0.4653	0.723	----	----
Jernv	----	----	----	----
Fotograf	-1.0744	0.552	----	----
Design	0.7243	0.629	----	----
Z	1.9435	0.028	1.8470	0.009
Nyreg	0.3360	0.726	----	----
R	-0.5414	0.446	----	----
Oms03	-0.8765	0.341	----	----
Oms310	-1.4337	0.114	----	----

A0	----	----	----	----
A1	1.7583	0.130	----	----
Komtjenest	0.4600	0.545	----	----
Komsentral	1.1876	0.190	1.2188	0.057
konstant	-4.4245	0.003	-4.6568	0.000
Antall obs	188		191	188
Log-likelihood	-38.7892		-43.2380	-42.9586
-2 log LR				8.3387
P-verdi redusert vs full modell				0.683

Estimatene i de reduserte modellene er basert på 191 observasjoner, mens LR testen er basert på de 188 observasjonene uten missings for den fulle modellen.

Noen variable måtte fjernes fra den fulle modellen på grunn av - i tillegg til kollinearitet for bransje og region som nevnt før - for liten informasjon i Y_2 (bare 14 av 191 verdier lik 1) som fører til at flere celler får "perfekt" prediksjon med estimert sannsynlighet lik 0 eller 1. Slike estimater er ikke spesielt nyttige. De vil føre til noen regresjonskoeffisienter lik pluss eller minus uendelig og de vil mangle standardavvik. Dette gjelder *ENK, Sor, Vest, Jernv* og *A0*.

Ved tolkning av tabell 4.7 over estimerte sannsynligheter se kommentar før tabell 4.4.

Tabell 4.7 **Sannsynligheter for avdekking av typen " endret merverdiavgift" , betinget av funn trinn 1 (** $Z = 1$**), for prediksjonsmodellen (fra tabell 4.6). Alle 4 bransjer.**

Funn trinn 1	Region	Kommune Mest sentral	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Funn	Midt	Ja	0.50 (2/4)	0.40	0.17
		Nei	0.00 (0/6)	0.16	0.06
	Ikke midt	Ja	0.18 (7/38)	0.17	0.10
		Nei	0.06 (2/35)	0.06	0.02
Ikke funn	Midt	Ja	0.00 (0/8)	0.09	0.03
		Nei	0.09 (2/22)	0.03	0.01
	Ikke midt	Ja	0.03 (1/39)	0.03	0.01
		Nei	0.00 (0/39)	0.01	0.00
		Total	0.07 (14/191)		

Når det gjelder avdekking av typen **"endret merverdiavgift relatert til avgiftsfeil"**, har screeningen på trinn 1 en klar effekt. Det synes også som Skatt Midt-Norge ligger noe høyere enn resten av landet for de 4 aktuelle bransjene når det gjelder forekomst av denne typen avvik (signifikansen for dette er imidlertid noe svak med p-verdi på 9%).

4.3 Sannsynligheten for avdekking av typen "påvist uteholdt omsetning" (indikator Y_3) gitt utfallet av screeningen på trinn 1.

Det var i alt 16 avdekkinger av denne typen av 191 i trinn-2-utvalget (8.4%)

Tabell 4.8 **Regresjonsresultater (logistisk regresjon) for avdekking av typen " påvist uteholdt omsetning" .**

(Basert på utskrift Ut 6-8 i appendiks A3)

Avhengig Y_3	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
AS	----	----	----	----	----	----
ENK	1.8902	0.175	----	----	----	----
Ost	0.2832	0.797	----	----	----	----
Sor	0.6099	0.591	----	----	----	----
Vest	-0.6482	0.649	----	----	----	----
Midt	0.9519	0.361	----	----	----	----
Nord	----	----	----	----	----	----
Snekker	0.0479	0.960	----	----	----	----
Jernv	1.3328	0.371	----	----	----	----
Fotograf	0.7781	0.455	----	----	----	----
Design	----	----	----	----	----	----
Z	0.5194	0.480	1.0038	0.083	----	----
Nyreg	-0.2196	0.813	----	----	----	----
R	-1.0392	0.095	-1.2600	0.022	-1.1545	0.038
Oms0_3	1.7058	0.169	----	----	----	----
Oms3_10	1.4072	0.256	----	----	----	----
Oms0_10	----	----	----	----	2.0272	0.055
A0	----	----	----	----	----	----
A1	----	----	----	----	----	----
Komtjenest	0.0991	0.897	----	----	----	----
Komsentral	1.1873	0.135	1.3300	0.030	1.2292	0.046
konstant	-6.3132	0.005	-2.9986	0.000	-4.0557	0.000
Antall obs	188		191	188	188	
Log-likelihood	-41.6936		-47.4641	-47.1124	-45.6025	
-2 log LR				10.8376	7.8178	
P-verdi redusert vs full modell				0.543	0.799	

Av samme grunn som beskrevet under avsnitt 4.2 måtte noen variable fjernes fra den fulle modellen. Spesielt måtte *AS* fjernes siden på grunn av "perfekt prediksjon" siden ingen av endringstilfellene var *AS*.

Tabell 4.9 **Sannsynligheter for avdekking av typen " påvist uteholdt omsetning" , betinget av funn trinn 1 (** $Z = 1$**), for prediksjonsmodell 1 (fra tabell 4.8). Alle 4 bransjer.**

Funn trinn 1	Ekstern Regnskaps-fører	Kommune Mest sentral	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Funn	Ja	Ja	0.11 (3/28)	0.13	0.06
		Nei	0.00 (0/27)	0.04	0.01
	Nei	Ja	0.43 (6/14)	0.34	0.19
		Nei	0.14 (2/14)	0.12	0.05
Ikke funn	Ja	Ja	0.06 (2/35)	0.05	0.02
		Nei	0.04 (2/47)	0.01	0.00
	Nei	Ja	0.08 (1/12)	0.16	0.07
		Nei	0.00 (0/14)	0.05	0.02
		Total	0.08(16/191)		

Tabell 4.10 **Sannsynligheter for avdekking av typen " påvist uteholdt omsetning" , betinget av funn trinn 1 (** $Z = 1$**), for prediksjonsmodell 2 (tabell 4.8). Alle 4 bransjer.**

Funn trinn 1	Ekstern regnskaps-fører	Kommune mest sentral	Omsetning under 1 mill.	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Funn	Ja	Ja	Ja	0.10 (2/21)	0.12	0.06
			Nei	0.17 (1/6)	0.02	0.00
		Nei	Ja	0.00 (0/18)	0.04	0.01
			Nei	0.00 (0/8)	0.01	0.00
	Nei	Ja	Ja	0.43 (6/14)	0.31	0.18
			Nei	- (0/0)	0.06	0.01
		Nei	Ja	0.17 (2/12)	0.12	0.05
			Nei	0.00 (0/2)	0.02	0.00
Ikke funn	Ja	Ja	Ja	0.11 (2/18)	0.12	0.06
			Nei	0.00 (0/17)	0.02	0.00
		Nei	Ja	0.10 (2/21)	0.04	0.01
			Nei	0.00 (0/26)	0.01	0.00
	Nei	Ja	Ja	0.14 (1/7)	0.31	0.18
			Nei	0.00 (0/4)	0.06	0.01
		Nei	Ja	0.00 (0/6)	0.12	0.05
			Nei	0.00 (0/8)	0.02	0.00
			Total	0.09 (16/188)		

Merk at sannsynlighetene i tabell 4.10 er de samme enten det var funn eller ikke på trinn 1. Det betyr at sannsynlighetene også kan tolkes som estimerte prevalens-sannsynligheter. Vi merker oss også at "unndratt omsetning" synes mest vanlig blant virksomheter med omsetning under 1 mill - særlig blant dem som ikke har ekstern regnskapsfører.

4.4 Sannsynligheten for avdekking av typen "endret nettoinntekt" (indikator Y_1), kontrollert for utfallet av screeningen på trinn 1.

Som beskrevet i avsnitt 4.1, ble det ikke funnet evidens i materialet for at screeningen på trinn 1 har hatt innflytelse på sannsynligheten for endringer av typen "endret nettoinntekt". Dette fører til at de estimerte prevalens-sannsynlighetene er de samme som angitt i tabell 4.4 og 4.5 - reprodusert her i tabell 4.11 og 4.12.

Tabell 4.11 **Sannsynligheter for avdekking av typen " endret nettoinntekt", kontrollert for funn på trinn 1. Basert på prediksjonsmodell 1 (tabell 4.1). Alle 4 bransjer.**

Enhetstype	Region Skatt Øst	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
ENK	Ja	0.28 (9/32)	0.300	0.187
	Nei	0.13 (13/99)	0.125	0.080
Ikke ENK	Ja	0.20 (1/5)	0.082	0.023
	Nei	0.02 (1/55)	0.029	0.009
	Total	0.13 (24/191)		

Tabell 4.12 **Sannsynligheter for avdekking av typen "endret nettoinntekt" , kontrollert for funn på trinn 1. Basert på prediksjonsmodell 2 (tabell 4.1). Alle 4 bransjer.**

Enhetstype	Kommune Mest sentral	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
ENK	Ja	0.23 (16/69)	0.230	0.158
	Nei	0.10 (6/62)	0.099	0.053
Ikke ENK	Ja	0.05 (1/20)	0.057	0.017
	Nei	0.03 (1/40)	0.022	0.006
	Total	0.13 (24/191)		

Sannsynligheten for avdekking av "endret nettoinntekt" synes å være betydelig høyere i Skatt øst enn andre regioner. Likeledes i mest sentrale kommuner sammenlignet med andre kommuner. Det er også betydelig forskjell mellom virksomhetstypene ENK og ikke-ENK der ENK har størst sannsynlighet. Kombinasjonen med høyest avdekking-sannsynlighet er ENK-virksomheter fra de mest sentrale kommunene, og den minst sannsynlige kombinasjonen er ikke-ENK-virksomheter (dvs. hovedsakelig AS) fra mindre sentrale kommuner. Hverken

Bransje, R, Nyreg, Komtjenest eller *Omsetning* synes å ha vesentlig betydning i henhold til evidensen i data.

Merk, imidlertid, at denne analysen for Y_1 vil bli erstattet med en annen analyse i avsnitt 4.7 som bygger på en ny modell for Y_1 - det vil si som en marginalfordeling i den simultane fordelingen for Y_1 og Y_3.

4.5 Sannsynligheten for avdekking av typen " endring av merverdiavgift relatert til avgiftsfeil" (indikator Y_2), kontrollert for utfallet av screeningen på trinn 1.

I dette tilfellet var det en tydelig tendens til at screeningen på trinn 1 hadde en effekt. Ved å kombinere tabell 3.1 og 4.6, finner vi vektoren av forklaringsvariable som synes å ha effekt på avdekkings-sannsynligheten.

$$U = (Ost, Midt, ENK, R, Komsentral, Komtjenest)$$

For de øvrige forklaringsvariablene som ikke er med i U, for eksempel *Alder* (via *Nyreg*) og antall ansatte (via *A0, A1*), er det ikke funnet evidens for innflytelse. Den kontrollerte sannsynligheten for $Y_2 = 1$ er gitt ved

$$P(Y_2 = 1 | U) = q \cdot p_1 + (1 - q) \cdot p_0$$

der

$$q = P(Z = 1 | Ost, ENK, R, KSminKTJ)$$
$$p_0 = P(Y_2 = 1 | Midt, Komsentral, Z = 0)$$
$$p_1 = P(Y_2 = 1 | Midt, Komsentral, Z = 1)$$

Vektoren U har 48 mulige verdier. Tabellen over sannsynlighetene er derfor splittet opp i to tabeller over virksomhetstype ENK og Ikke-ENK (som hovedsakelig (87%) består av AS). Utviklingen av konfidensgrensene er litt mer komplisert i dette tilfellet og følger av metoden skissert i matematisk appendikser i HR og SR. De er utviklet først for den lineære prediktoren på logistisk skala (basert på asymptotisk teori) og deretter transformert til sannsynlighetsskalaen.

Når det gjelder tolking av tabellene, se kommentaren rett før tabell 4.4.

Tabell 4.13a – For ENK-virksomheter:

Sannsynligheter for avdekking av typen " endret merverdiavgift" (Y_2), kontrollert for funn på trinn 1. Alle 4 bransjer. Basert på prediksjonsmodeller fra tabell 3.1 og 4.6.

Ekstern regnskaps-fører	Region	Kommune mest sentral	Kommune hovedsakelig tjenesteytende	Rel. frekvens	Sanns. het	Ensidig nedre 95% konf. grense
Ja	Skatt øst	Ja	Ja	0.13 (2/16)	0.090	0.051
			Nei	0.00 (0/4)	0.072	0.029
		Nei	Ja	- (0/0)	0.036	0.019
			Nei	0.00 (0/2)	0.030	0.012
	Skatt midt	Ja	Ja	0.00 (0/3)	0.153	0.064
			Nei	0.00 (0/3)	0.131	0.050
		Nei	Ja	0.00 (0/1)	0.069	0.026
			Nei	0.22 (2/9)	0.056	0.020
	Andre	Ja	Ja	0.17 (1/6)	0.057	0.030
			Nei	0.00 (0/16)	0.047	0.017
		Nei	Ja	0.10 (1/10)	0.023	0.012
			Nei	0.04 (1/24)	0.018	0.007
Nei	Skatt øst	Ja	Ja	0.22 (2/9)	0.107	0.061
			Nei	0.00 (0/1)	0.087	0.036
		Nei	Ja	- (0/0)	0.042	0.022
			Nei	- (0/0)	0.035	0.014
	Skatt midt	Ja	Ja	1.00 (2/2)	0.179	0.079
			Nei	- (0/0)	0.149	0.060
		Nei	Ja	0.00 (0/2)	0.084	0.031
			Nei	0.00 (0/2)	0.067	0.025
	Andre	Ja	Ja	0.25 (1/4)	0.070	0.038
			Nei	0.00 (0/5)	0.056	0.021
		Nei	Ja	0.00 (0/5)	0.029	0.015
			Nei	0.00 (0/7)	0.023	0.009
			Total	0.09 (12/131)		

Tabell 4.13b – For Ikke-ENK virksomheter :

Sannsynligheter for avdekking av typen " endret merverdiavgift" (Y_2), kontrollert for funn på trinn 1. Alle 4 bransjer. Basert på prediksjonsmodeller fra tabell 3.1 og 4.6.

Ekstern regnskaps-fører	Region	Kommune mest sentral	Kommune hovedsakelig tjenesteytende	Rel. frekvens	Sanns. het	Ensidig nedre 95% konf. grense
Ja	Skatt øst	Ja	Ja	0.33 (1/3)	0.053	0.026
			Nei	1.00 (1/1)	0.044	0.015
		Nei	Ja	- (0/0)	0.021	0.010
			Nei	- (0/0)	0.017	0.006
	Skatt midt	Ja	Ja	0.00 (0/2)	0.112	0.040
			Nei	- (0/0)	0.105	0.034
		Nei	Ja	0.00 (0/6)	0.043	0.015
			Nei	0.00 (0/5)	0.037	0.012

	Andre	Ja	Ja	0.00 (0/6)	0.039	0.016
			Nei	0.00 (0/3)	0.036	0.011
		Nei	Ja	0.00 (0/9)	0.014	0.006
			Nei	0.00 (0/8)	0.012	0.004
Nei	Skatt øst	Ja	Ja	- (0/0)	0.063	0.033
			Nei	0.00 (0/1)	0.051	0.019
		Nei	Ja	- (0/0)	0.026	0.013
			Nei	- (0/0)	0.020	0.008
	Skatt midt	Ja	Ja	0.00 (0/1)	0.121	0.045
			Nei	0.00 (0/1)	0.110	0.038
		Nei	Ja	0.00 (0/1)	0.049	0.018
			Nei	0.00 (0/2)	0.042	0.014
	Andre	Ja	Ja	- (0/0)	0.043	0.019
			Nei	0.00 (0/2)	0.038	0.012
		Nei	Ja	0.00 (0/3)	0.016	0.008
			Nei	0.00 (0/6)	0.014	0.005
			Total	0.03 (2/60)		

Vi ser at kun 2 av de 60 Ikke-ENK-virksomhetene (hvorav 52 AS) i materialet hadde avdekking av typen " endret merverdiavgift". Dette er såpass tynt at de estimerte sannsynlighetene bør tas med en klype salt – de er uttrykk for tendenser som først og fremst gjør seg gjeldende for ENK-virksomheter.

For øvrig ses at de estimerte sannsynlighetene er høyest i Skatt Midt-Norge, noe lavere i Skatt Øst og lavest i Sør, Vest og Nord. Virksomheter uten ekstern regnskapsfører ligger litt høyere i sannsynlighet (gjennomsnittlig rundt 20% høyere) enn de med. De kombinasjonene med høyest sannsynlighet i alle regioner er ENK-virksomheter uten ekstern regnskapsfører fra de mest sentrale kommunene. Forskjellen mellom hovedsakelig tjenesteytende kommuner og andre synes ubetydelig.

4.6 Sannsynligheten for avdekking av typen "påvist uteholdt omsetning" (indikator Y_3), kontrollert for utfallet av screeningen på trinn 1.

Utviklingen i avsnitt 4.3 ledet til to alternative prediksjonsmodeller som data ikke har informasjon nok til å skille mellom. I prediksjonsmodell 1 i tabell 4.8 inngår *Z* som forklaringsvariabel som nødvendiggjør kontroll for funn på trinn 1. I prediksjonsmodell 2 inngår ikke *Z* som forklaringsvariabel slik at kontroll for funn på trinn 1 ikke er nødvendig.

4.6.1 Prevalens-sannsynligheter for prediksjonsmodell 1 (tabell 4.8):

I dette tilfellet var det noe, men ikke sterk, evidens for at screeningen på trinn 1 hadde en effekt. Ved å kombinere tabell 3.1 og 4.8, finner vi vektoren av forklaringsvariable som ifølge prediksjonsmodell 1 synes å ha effekt på avdekking-sannsynligheten.

$$U = (Ost, R, ENK, Komsentral, Komtjenest)$$

For de øvrige forklaringsvariablene som ikke er med i U, for eksempel *Alder* (via *Nyreg*), antall ansatte (via *A0, A1*) og *Omsetning* (via *Oms0_3* og *Oms3_10* - se imidlertid prediksjonsmodell 2 under), er det ikke funnet evidens for innflytelse. Den kontrollerte sannsynligheten for $Y_3 = 1$ er gitt ved

$$P(Y_3 = 1 | U) = q \cdot p_1 + (1 - q) \cdot p_0$$

der

$$q = P(Z = 1 | Ost, R, ENK, KS \min KTJ)$$
$$p_0 = P(Y_3 = 1 | R, Komsentral, Z = 0)$$
$$p_1 = P(Y_3 = 1 | R, Komsentral, Z = 1)$$

Vektoren U har 48 mulige verdier. Tabellen over sannsynlighetene er derfor splittet opp i to tabeller for virksomhetstypene ENK og ikke-ENK (hvorav 87% er AS). Utviklingen av konfidensgrensene følger av metoden skissert i HR

Når det gjelder tolking av tabellene, se kommentaren rett før tabell 4.4.

Tabell 4.14a – ENK-virksomheter:

Sannsynligheter for avdekking av typen "påvist uteholdt omsetning" (Y_3), kontrollert for funn på trinn 1. Alle 4 bransjer. Basert på prediksjonsmodell 1 fra tabell 3.1 og 4.8.

Skatt Øst	Ekstern regnskaps-fører	Kommune mest sentral	Kommune hovedsakelig tjenesteytende	Rel. frekvens	Sanns. het	Ensidig nedre 95% konf. grense
Ja	Ja	Ja	Ja	0.13 (2/16)	0.083	0.043
			Nei	0.00 (0/4)	0.073	0.028
		Nei	Ja	- (0/0)	0.027	0.013
			Nei	0.00 (0/2)	0.024	0.009
	Nei	Ja	Ja	0.33 (3/9)	0.258	0.149
			Nei	0.00 (0/1)	0.232	0.108
		Nei	Ja	- (0/0)	0.097	0.050
			Nei	- (0/0)	0.087	0.036
Nei	Ja	Ja	Ja	0.33 (3/9)	0.065	0.032
			Nei	0.00 (0/19)	0.060	0.022
		Nei	Ja	0.00 (0/11)	0.021	0.010
			Nei	0.06 (2/33)	0.018	0.007
	Nei	Ja	Ja	0.17 (1/6)	0.209	0.113
			Nei	0.60 (3/5)	0.191	0.081
		Nei	Ja	0.00 (0/7)	0.077	0.040
			Nei	0.11 (1/9)	0.068	0.027
			Total	0.11 (15/131)		

Tabell 4.14b – Ikke-ENK-virksomheter:

Sannsynligheter for avdekking av typen "påvist uteholdt omsetning" (Y_3), kontrollert for funn på trinn 1. Alle 4 bransjer. Basert på prediksjonsmodell 1 fra tabell 3.1 og 4.8.

Skatt Øst	Ekstern regnskaps-fører	Kommune mest sentral	Kommune hovedsakelig tjenesteytende	Rel. frekvens	Sanns. het	Ensidig nedre 95% konf. grense
Ja	Ja	Ja	Ja	0.00 (0/3)	0.063	0.030
			Nei	0.00 (0/1)	0.058	0.021
		Nei	Ja	- (0/0)	0.020	0.010
			Nei	- (0/0)	0.018	0.006
	Nei	Ja	Ja	- (0/0)	0.200	0.105
			Nei	0.00 (0/1)	0.185	0.076
		Nei	Ja	- (0/0)	0.073	0.037
			Nei	- (0/0)	0.064	0.025
Nei	Ja	Ja	Ja	0.00 (0/8)	0.055	0.024
			Nei	0.00 (0/3)	0.053	0.018
		Nei	Ja	0.00 (0/15)	0.016	0.007
			Nei	0.00 (0/13)	0.015	0.005
	Nei	Ja	Ja	0.00 (0/1)	0.174	0.083
			Nei	0.00 (0/3)	0.168	0.064
		Nei	Ja	0.00 (0/4)	0.058	0.027
			Nei	0.13 (1/8)	0.054	0.020
			Total	0.02 (1/60)		

Vi ser at kun 1 av de 60 Ikke-ENK-virksomhetene (hvorav 52 AS) i materialet hadde avdekking av typen " uteholdt omsetning". Dette er såpass tynt at de estimerte sannsynlighetene bør tas med en klype salt – de er uttrykk for tendenser som først og fremst gjør seg gjeldende for ENK-virksomheter.

Når det gjelder ENK-virksomheter ses at sannsynlighetene for Skatt øst ligger her litt over de øvrige. Den klareste kontrasten i sannsynlighet er mellom virksomheter med og uten ekstern regnskapsfører. Det er forholdsvis ubetydelige forskjeller mellom virksomheter fra hovedsakelig tjenesteytende kommuner og andre.

4.6.2 Prevalens-sannsynligheter for prediksjonsmodell 2 (tabell 4.8):

I denne prediksjonsmodellen inngår ikke *Z*. Det er altså ikke evidens i data for at *Z* har betydning utover *R, Komsentral* og "*omsetning under 1 mill*". Dette fører til at de estimerte prevalens-sannsynlighetene er de samme som angitt i tabell 4.10 - reprodusert her i tabell 4.15.

Tabell 4.15 Sannsynligheter for avdekking av typen "påvist uteholdt omsetning" (Y_3), kontrollert for funn på trinn 1. Alle 4 bransjer og alle regioner. Basert på prediksjonsmodell 2 fra tabell 4.8.

Ekstern regnskaps-fører	Kommune mest sentral	Omsetning under 1 mill.	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Ja	Ja	Ja	0.10 (4/39)	0.124	0.064
		Nei	0.04 (1/23)	0.018	0.003
	Nei	Ja	0.05 (2/39)	0.040	0.015
		Nei	0.00 (0/34)	0.005	0.001
Nei	Ja	Ja	0.33 (7/21)	0.310	0.181
		Nei	0.00 (0/4)	0.056	0.010
	Nei	Ja	0.11 (2/18)	0.116	0.048
		Nei	0.00 (0/10)	0.017	0.003
		Total	0.08 (16/188)		

Vi merker oss at variabelen "*omsetning under 1 mill*" har en klar effekt. For det første synes sannsynligheten for avdekking å øke med en faktor på ca 6-7 for virksomheter med omsetning under 1 mill. i forhold til virksomheter med omsetning over 1 mill. I tillegg synes variabelen å fjerne effekten av funn på trinn 1 (som gjør det unødvendig å kontrollere for screeningen på trinn 1), og dessuten effektene av *Region og Komtjenest*, som inngår, direkte eller indirekte, ved bruk av prediksjonsmodell 1.

Om vi sammenligner prediksjonsmodellene 1 og 2 med informasjonskriteriene AIC og BIC (nærmere beskrevet i HR), har prediksjonsmodell 1 AIC = 102.9 og BIC = 115.9, mens prediksjonsmodell 2 fikk verdiene AIC = 99.2 og BIC = 112.2, som klart styrker prediksjonsmodell 2 i forhold til 1. Det er interessant at denne analysen toner ned betydningen av hovedsakelig tjenesteytende kommune, mens ekstern regnskapsfører og omsetning kommer klarere fram. Det er også interessant at effekten av disse to variablene synes å forsvinne (jfr. tabell 4.1) når vi slår sammen responsen "uteholdt omsetning" med andre årsaker til responsen "endret nettoinntekt".

4.7 Modellering av simultanfordelingen for "endret nettoinntekt" (Y_1) og "påvist uteholdt omsetning" (Y_3), kontrollert for utfallet av screeningen på trinn 1.

Som nevnt i innledningen er det en sterk sammenheng mellom Y_1 og Y_3 ved at "påvist uteholdt omsetning" impliserer "endret nettoinntekt". Dermed må vi ha $Y_3 \le Y_1$, som impliserer

(4.7.1) $$P(Y_3 = 1 | U) \le P(Y_1 = 1 | U)$$

der U er en vektor av kovariater. Nå er Y_1 og Y_3 analysert uavhengig av hverandre i avsnitt 4.4 og 4.6, og vi har ingen garanti for at relasjon (4.7.1) gjelder for de estimerte sannsynlighetene. Om vi sammenligner de estimerte sannsynlighetene for felles kovariater,

som er gjort i tabell 4.16, finner vi faktisk flere selvmotsigelser - det vil si verdier som strider mot (4.7.1).

Tabell 4.16 Sammenligning av estimerte sannsynligheter for "endret nettoinntekt" (Y_1) og "påvist uteholdt omsetning" (Y_3) ut fra tabell 4.11 og tabell 4.15.

ENK	Ekstern regnskaps-fører	Kommune mest sentral	Omsetning under 1 mill.	Tabell 4.11 Sanns.het for "endret netto-inntekt"	Tabell 4.15 Sanns.het for "påvist uteholdt omsetning"	
Ja	Ja	Ja	Ja	0.230	0.124	
			Nei	0.230	0.018	
		Nei	Ja	0.099	0.040	
			Nei	0.099	0.005	
	Nei	Ja	Ja	0.230	0.310	selvmots
			Nei	0.230	0.056	
		Nei	Ja	0.099	0.116	selvmots
			Nei	0.099	0.017	
Nei	Ja	Ja	Ja	0.057	0.124	selvmots
			Nei	0.057	0.018	
		Nei	Ja	0.022	0.040	selvmots
			Nei	0.022	0.005	
	Nei	Ja	Ja	0.057	0.310	selvmots
			Nei	0.057	0.056	
		Nei	Ja	0.022	0.116	selvmots
			Nei	0.022	0.017	

En mulighet for å unngå slike selvmotsigelser, er å modellere den simultane fordelingen for Y_1 og Y_3. Dette er gjennomført i Appendiks 1 på grunn av noe mer teknisk pregete detaljer. Tabell 4.17 og 4.18 er også flyttet til appendiks 1.

Analysen i appendiks 1 styrker inntrykket at funn på trinn 1 (*Z*) heller ikke for denne fordelingen synes å ha betydning. Sammen med analysen i avsnitt 4.6 har vi således ikke funnet evidens for at *Z* har betydning for simultanfordelingen til Y_1 og Y_3. Jeg antar dermed dette som forutsetning nedenfor - nemlig at *Z* ikke har betydning for simultanfordelingen til Y_1 og Y_3 - som forenkler analysen en del.

Tabell 4.19 **Sannsynligheter for avdekking av typen "endret nettoinntekt" , kontrollert for funn på trinn 1 og for $Y_3 \le Y_1$. Basert på tabell 4.17 i appendiks 1 og 4.8 (prediksjonsmodell 2). Alle 4 bransjer.**

Skatt Øst	Ekstern regnskaps-fører	Kommune mest sentral	Omsetning under 1 mill.	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Ja	Ja	Ja	Ja	0.20 (3/15)	0.261	0.164
			Nei	0.50 (4/8)	0.172	0.089
		Nei	Ja	0.00 (0/1)	0.190	0.105
			Nei	0.00 (0/1)	0.161	0.081
	Nei	Ja	Ja	0.33 (3/9)	0.418	0.282
			Nei	0.00 (0/1)	0.203	0.104
		Nei	Ja	-- (0/0)	0.254	0.150
			Nei	-- (0/0)	0.171	0.088
Nei	Ja	Ja	Ja	0.13 (3/24)	0.143	0.080
			Nei	0.00 (0/15)	0.039	0.015
		Nei	Ja	0.08 (3/38)	0.060	0.029
			Nei	0.06 (2/33)	0.026	0.011
	Nei	Ja	Ja	0.33 (4/12)	0.325	0.196
			Nei	0.00 (0/3)	0.076	0.021
		Nei	Ja	0.11 (2/18)	0.135	0.063
			Nei	0.00 (0/10)	0.038	0.014
			Total	0.13 (24/188)		

Det kan nå ha interesse å sammenligne $P(Y_1 = 1|U)$ fra den kombinerte analysen av Y_1 og Y_3 fra tabell 4.19 med den separate analysen av Y_1 fra tabell 4.11. Dette er gjort i tabell 4.20a og b, splittet opp på Skatt Øst/ ikke-Øst.

Tabell 4.20a - Skatt Øst

Sammenligning av estimert $P(Y_1 = 1|U)$ ved separat analyse av Y_1 (tabell 4.11), analyse av Y_3 (tabell 4.15) og $P(Y_1 = 1|U)$ ved kombinert analyse av Y_1 og Y_3 (tabell 4.19).

				Separat analyse		Kombinert analyse
ENK	Ekstern regnskaps-fører	Kommune mest sentral	Omsetning under 1 mill.	Tabell 4.11 Sanns.het for "endret netto-inntekt"	Tabell 4.15 Sanns.het for "påvist uteholdt omsetning"	Tabell 4.19 Sanns.het for "endret netto-inntekt"
Ja	Ja	Ja	Ja	0.230	0.124	0.261
			Nei	0.230	0.018	0.172
		Nei	Ja	0.099	0.040	0.190

			Nei	0.099	0.005	0.161
	Nei	Ja	Ja	0.230	0.310	0.418
			Nei	0.230	0.056	0.203
		Nei	Ja	0.099	0.116	0.254
			Nei	0.099	0.017	0.171
Nei	Ja	Ja	Ja	0.057	0.124	0.261
			Nei	0.057	0.018	0.172
		Nei	Ja	0.022	0.040	0.190
			Nei	0.022	0.005	0.161
	Nei	Ja	Ja	0.057	0.310	0.418
			Nei	0.057	0.056	0.203
		Nei	Ja	0.022	0.116	0.254
			Nei	0.022	0.017	0.171

Tabell 4.20b - Skatt ikke-Øst

Sammenligning av estimert $P(Y_1 = 1 | U)$ ved separat analyse av Y_1 (tabell 4.11), analyse av Y_3 (tabell 4.15) og $P(Y_1 = 1 | U)$ ved kombinert analyse av Y_1 og Y_3 (tabell 4.19).

				Separat analyse		Kombinert analyse
ENK	Ekstern regnskaps-fører	Kommune mest sentral	Omsetning under 1 mill.	Tabell 4.11 Sanns.het for "endret netto-inntekt"	Tabell 4.15 Sanns.het for "påvist uteholdt omsetning"	Tabell 4.19 Sanns.het for "endret netto-inntekt"
Ja	Ja	Ja	Ja	0.230	0.124	0.143
			Nei	0.230	0.018	0.039
		Nei	Ja	0.099	0.040	0.060
			Nei	0.099	0.005	0.026
	Nei	Ja	Ja	0.230	0.310	0.325
			Nei	0.230	0.056	0.076
		Nei	Ja	0.099	0.116	0.135
			Nei	0.099	0.017	0.038
Nei	Ja	Ja	Ja	0.057	0.124	0.143
			Nei	0.057	0.018	0.039
		Nei	Ja	0.022	0.040	0.060
			Nei	0.022	0.005	0.026
	Nei	Ja	Ja	0.057	0.310	0.325
			Nei	0.057	0.056	0.076
		Nei	Ja	0.022	0.116	0.135
			Nei	0.022	0.017	0.038

Så hvilke sannsynligheter for Y_1 skal man tro på - de fra den separate eller de fra den kombinerte analysen? Til tross for "selvmotsigelsene" i tabell 4.16, er den separate analysen mulig - med spesiell vekt på de nedre konfidensgrensene. Selvmotsigelsene kan oppfattes som et uttrykk for usikkerheten - dels den som er aktuell gitt prediksjonsmodellen og dels den som

er knyttet til valget av prediksjonsmodell. Dette er et typisk dilemma ved begrensete datamaterialer der informasjonsmengden ikke er stor nok til å diskriminere effektivt mellom flere alternative modeller og analyser. En svakhet ved den "riktige" kombinerte analysen, for eksempel, er at det inngår en logistisk regresjon av bare 8 $(Y_1 = 1)$ - hendelser blant 175, som er mindre informativ enn i den separate analysen som omfatter 24 $(Y_1 = 1)$- hendelser blant 191.

4.8 Flere typer av "endret nettoinntekt"

Det er andre måter å unngå "selvmotsigelsene" i tabell 4.16 - for eksempel ved å skille mellom når beløpet for "endret nettoinntekt" er større enn "uteholdt omsetning" ($X_1 > X_3$) og når de er like. Da vil $X_1 - X_3$ representere beløpet for andre typer av "endret nettoinntekt" enn "uteholdt omsetning". La Y_{1a} være indikatoren for andre typer av "endret nettoinntekt".

(4.8.1) $$Y_{1a} = \begin{cases} 1 & \text{hvis } X_1 > X_3 \\ 0 & \text{ellers} \end{cases}$$

Da kan Y_1 uttrykkes ved

$$Y_1 = \text{maks}\{Y_{1a}, Y_3\}$$

Tabell 4.21 **Simultan frekvenstabell for Y_{1a} og Y_3**

Uteholdt oms. (Y_3)	Y_{1a}		
	0	1	Sum
0	167	8	175
1	10	6	16
Sum	177	14	191

Ved denne kategoriseringen blir hendelsen $(Y_1 = 1)$ splittet opp i tre kategorier, $(Y_{1a}, Y_3) = (0, 1),\ (1, 0)$ eller $(1, 1)$ med frekvenser 8, 10 og 6 h.h.v. Simultanfordelingen for (Y_{1a}, Y_3) som funksjon av et passende utvalg av kovariater, kan for eksempel analyseres ved en multinomisk logistisk analyse - noe som ikke er gjennomført her. Jeg vil nøye meg med en logistisk analyse av Y_{1a} i tråd med metodikken ovenfor og oppsummert i tabell 4.22 og 4.23.

Tabell 4.22 Regresjonsresultater (logistisk regresjon) for avdekking av typen "endret nettoinntekt" av andre typer enn "påvist uteholdt omsetning".

(Basert på utskrift Ut13 og Ut14 i appendiks A5)

Avhengig Y_{1a}	Full modell		Prediksjonsmodell	
	Koeff.	p-verdi	Koeff.	p-verdi
AS	-----	-----	-----	-----
ENK	2.7034	0.085	2.2691	0.043
Ost	1.6797	0.134	1.4215	0.017
Sor	1.0903	0.313	-----	-----
Vest	-0.0297	0.982	-----	-----
Midt	-----	-----	-----	-----
Nord	-----	-----	-----	-----
Snekker	0.7790	0.451	-----	-----
Jernv	1.3441	0.356	-----	-----
Fotograf	1.0718	0.369	-----	-----
Design	-----	-----	-----	-----
Z	-0.5224	0.503	-----	-----
Nyreg	-1.2041	0.323	-----	-----
R	0.5549	0.505	-----	-----
Oms0_3	0.2981	0.757	-----	-----
Oms3_10	-1.0681	0.222	-----	-----
Oms0_10	-----	-----	-1.1398	0.071
A0	-0.6031	0.715	-----	-----
A1	-0.2945	0.866	-----	-----
Komtjenest	0.5306	0.520	-----	-----
Komsentral	0.7856	0.396	-----	-----
konstant	-6.4113	0.003	-4.1113	0.000
Antall obs	188		188	
Log-likelihood	-38.6536		-42.6647	
-2 log LR			8.0221	
P-verdi redusert vs full modell			0.842	

Tabell 4.23 Sannsynligheter for avdekking av typen "endret nettoinntekt" av andre typer enn "påvist uteholdt omsetning". Basert på tabell 4.22.

Skatt Øst	ENK	Omsetning under 1 mill.	Rel. frekvens	Sanns.het	Ensidig nedre 95% konf. grense
Ja	Ja	Ja	0.13 (3/23)	0.105	0.038
		Nei	0.43 (3/7)	0.440	0.197
	Nei	Ja	0.00 (0/2)	0.018	0.002
		Nei	0.33 (1/3)	0.109	0.018
Nei	Ja	Ja	0.06 (5/81)	0.013	0.004
		Nei	0.12 (2/17)	0.083	0.026
	Nei	Ja	0.00 (0/11)	0.002	0.000
		Nei	0.00 (0/44)	0.014	0.002
		Total	0.07 (14/188)		

Vi merker oss at de største sannsynlighetene finnes blant virksomheter med omsetning over 1 mill., og at Skatt øst også her synes å ligge over resten av landet.

5. Estimering av forventet endringsbeløp gitt endring

5.1 Separat analyse av X_1 (beløp for "endret nettoinntekt")

Datagrunnlaget for X_1 er 24 observasjoner. I tabell 5.1 har jeg foretatt den innledende analysen av X_1 basert på generalisert lineær modellering (GLM), som beskrevet i HR og SR, med gammafordeling og log linkfunksjon.

I tabellen inngår to nye variable, *OstVest* og *NyregR*. *OstVest* er en indikator for Skatt Øst og Vest slått sammen. Denne variabelen behøves ikke i den fulle modellen siden den kan genereres ved en restriksjon på koeffisientene for *Ost* og *Vest*. *NyregR* er en samspillsvariabel definert som produktet av *R* (ekstern regnskapsfører) og *Nyreg* (alder høyst 3 år). Den viser seg å fungere bedre enn *Nyreg*, og erstatter *Nyreg* siden bruk av begge skaper kollinearitetsproblemer.

I avsnitt 5.1.1 blir det påvist at den valgte log-lineære modellen med gammafordeling kanskje ikke er helt realistisk i og med at den ikke synes å håndtere tendensen til ekstreme observasjoner tilfredsstillende. Det er derfor også beregnet såkalte "robuste" p-verdier (oppgitt i parentes) i tabellene 5.1 og 5.2. Robuste p-verdier er beregnet på grunnlag av "robuste" standardavvik som i litteraturen går under navnet Huber/White's "sandwich"-type estimatorer – og som til en viss grad kompenserer for bruk av feilaktig likelihoodfunksjon.

Tabell 5.1 **Regresjonsresultater (GLM, gamma og log link) for endringsbeløp av typen "endret nettoinntekt" gitt endring og funn/ikke-funn på trinn 1. Prediksjonsmodell 1. (Robuste p-verdier i parentes)**

(Basert på utskrift Ut15 og Ut16 i appendiks A6)

Avhengig X_1	Full modell		Prediksjonsmodell 1	
	Koeff.	p-verdi	Koeff.	p-verdi
AS	-----	-----	-----	-----
ENK	2.8951	0.187	2.3547	0.000 (0.000)
Ost	-0.8627	0.400	-----	-----
Sor	-0.1342	0.929	-----	-----
Vest	-0.9410	0.492	-----	-----
Midt	0.2168	0.875	-----	-----
Nord	-----	-----	-----	-----

OstVest	-----	-----	-1.1143	0.014 (0.027)
Snekker	-1.6169	0.312	-----	-----
Jernv	-----	-----	-----	-----
Fotograf	-1.2639	0.336	-----	-----
Design	-1.8632	0.287	-----	-----
Z	1.6930	0.034	1.1581	0.005 (0.020)
R	3.0061	0.000	2.0780	0.000 (0.005)
NyregR	-2.3567	0.018	-2.0633	0.002 (0.007)
Oms0_3	-----	-----	-----	-----
Oms3_10	-----	-----	-----	-----
Oms0_10	1.4474	0.057	1.3888	0.007 (0.008)
A0	-2.6398	0.148	-2.9581	0.000 (0.000)
A1	-0.8473	0.646	-----	-----
Komtjenest	-0.6001	0.618	-----	-----
Komsentral	0.6745	0.626	-----	-----
konstant	9.0263	0.000	9.6356	0.000 (0.000)
Antall obs	24		24	
Log-likelihood	-293.2170		-294.7976	
-2 log LR			3.1612	
P-verdi redusert vs full modell			0.9998	
AIC	25.85		25.23	
BIC	-10.40		-35.85	

Det første vi legger merke til, er at screeningen på trinn 1 (*Z*) synes å ha en klar effekt som, om den er reell, bør kontrolleres for. Det er grunn til å tro at avhengigheten av *Z*, først og fremst skyldes to ekstreme observasjoner, 500 000 og 685 787 (de 22 øvrige er under 220 000). Hvis vi gjennomfører samme analyse uten disse to, forsvinner *Z*'s innflytelse helt, og aktuelle prediksjonsmodeller får høyst to forklaringsvariable. Virksomhetene til de to ekstreme observasjonene har dessuten forskjellig verdi for *Z*. Det at *Z*'s betydning først og fremst synes knyttet til effekten av de to store observasjonene gir grunnlag for å tvile på at *Z*'s betydning er reell. Dette er noe utdypet i avsnitt 5.1.1, der det bl.a. fremgår at prediksjonsmodell 1 også har visse andre uheldige egenskaper som begrenser dens nytte.

Det er også påfallende at den reduserte modellen inneholder relativt mange forklaringsvariable. Det er grunn til å tro at dette er et utslag av overtilpasning av prediksjonsmodell 1 (jfr. diskusjonen av X_2 i avsnitt 5.2) som innebærer reduserte prediksjonsegenskaqper. Den ekstremt høye p-verdien ved testing av prediksjonsmodellen mot den fulle modellen er et symptom på overtilpasning. Jeg har derfor foreslått to alternative prediksjonsmodeller i tabell 5.2 med bedre prediksjonsegenskaper enn prediksjonsmodell 1, og der *Z* ikke har betydning (begrunnet i avsnitt 5.1.1).

Tabell 5.2 **Regresjonsresultater (GLM, gamma og log link) for endringsbeløp av typen "endret nettoinntekt" gitt endring og funn/ikke-funn på trinn 1. Prediksjonsmodell 2 og 3. (Robuste p-verdier i parentes)**

(Basert på utskrift Ut17 og Ut18 i appendiks A6)

Avhengig X_1	Full modell		Prediksjonsmodell 2		Prediksjonsmodell 3	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
AS	-----	-----	-----	-----	-----	-----
ENK	2.8951	0.187	1.6378	0.058 (0.000)	2.3478	0.029 (0.000)
Ost	-0.8627	0.400	-----	-----	-----	-----
Sor	-0.1342	0.929	-----	-----	-----	-----
Vest	-0.9410	0.492	-----	-----	-----	-----
Midt	0.2168	0.875	-----	-----	-----	-----
Nord	-----	-----	-----	-----	-----	-----
OstVest	-----	-----	-----	-----	-----	-----
Snekker	-1.6169	0.312	-----	-----	-----	-----
Jernv	-----	-----	-----	-----	-----	-----
Fotograf	-1.2639	0.336	-----	-----	-----	-----
Design	-1.8632	0.287	-----	-----	-----	-----
Z	1.6930	0.034	-----	-----	-----	-----
R	3.0061	0.000	0.8284	0.092 (0.058)	-----	-----
NyregR	-2.3567	0.018	-----	-----	-----	-----
Oms0_3	-----	-----	-----	-----	-----	-----
Oms3_10	-----	-----	-----	-----	-----	-----
Oms0_10	1.4474	0.057	-----	-----	-----	-----
A0	-2.6398	0.148	-----	-----	-1.3819	0.198 (0.000)
A1	-0.8473	0.646	-----	-----	-----	-----
Komtjenest	-0.6001	0.618	-----	-----	-----	-----
Komsentral	0.6745	0.626	-----	-----	-----	-----
konstant	9.0263	0.000	9.4844	0.000 (0.000)	10.6034	0.000 (0.000)
Antall obs	24		24		24	
Log-likelihood	-293.2170		-300.0817		-299.7328	
-2 log LR			13.7295		13.0316	
P-verdi redusert vs full modell			0.470		0.524	
AIC	25.85		25.26		25.23	
BIC	-10.40		-41.17		-41.87	

Noen prediksjoner basert på PM 2 og PM3 er gitt i avsnitt 5.1.2.

5.1.1 Utdypende diskusjon av prediksjonsmodell 1, 2 og 3 for "endret nettoinntekt".

Det at den viktige variabelen *Z* (som bestemmer skjevheter som følge av screeningen på trinn 1) dukker opp som tilsynelatende betydningsfull i prediksjonsmodell 1 (PM1) gjør at man bør ha gode argumenter for velge en prediksjonsmodell uten *Z*. Mitt grunnlag for å velge en prediksjonsmodell uten *Z* bygger på en mistanke om det at *Z* kommer med i PM1 først og fremst skyldes tilfeldigheter ved at de to største observasjonene X_1 har forskjellig verdi av *Z* kombinert med det at den største observasjonen synes å ha en utilbørlig stor innflytelse på estimeringen. Den følgende argumentasjonen er en utdyping av dette.

Vi ønsker å estimere forventet endringsbeløp av typen "endret nettoinntekt" gitt endring, symbolsk $E(X_1 \mid U, Y_1 = 1)$, der *U* er en vektor av kovariater bestemt av de valgte prediksjonsmodellene. Forventet endring gitt bare *U*, $E(X_1 \mid U)$, representerer en sammenblanding av tilfeller der $X_1 = 0$ og der $X_1 > 0$, og kan beregnes (jfr. matematisk appendiks i SR) ved $E(X_1 \mid U) = E(X_1 \mid U, Y_1 = 1) \cdot P(Y_1 = 1 \mid U)$. Den har mindre tolkningsmessig interesse og er først og fremst nyttig ved aggregering til totaltall over grupper av strata, som er gjort rede for i HR. Den type aggregering er ikke foretatt her.

Prediksjonsmodell 1 (PM1) i tabell 5.1 gir grunnlag for estimering av $\mu_z(U_1) = E(X_1 \mid U_1, Y_1 = 1, Z = z)$, der *Z* er med, og der U_1 består av kovariatene

$$U_1 = (ENK, \text{ØstVest}, R, \text{NyregR}, A_0, Oms_{0-10})$$

For å kunne aggregere ut *Z*, trenger vi også $q(U_2) = P(Z = 1 \mid U_2)$, der U_2 består av kovariatene

$$U_2 = (\text{Øst}, R, ENK, KS \min Ktj)$$

Tilsvarende som i appendiks A1.1 i SR finner vi

$$E(X_1 \mid U, Y_1 = 1) = \mu_0(U_1)\left(1 + \left(e^{\beta_z} - 1\right) q(U_2)\right) \tag{5.1.1}$$

der β_z er regresjonskoeffisienten for *Z* i den generalisert lineære analysen for X_1, og der den totale kovariat-vektoren, *U*, består av alle variable som inngår i U_1 og U_2.

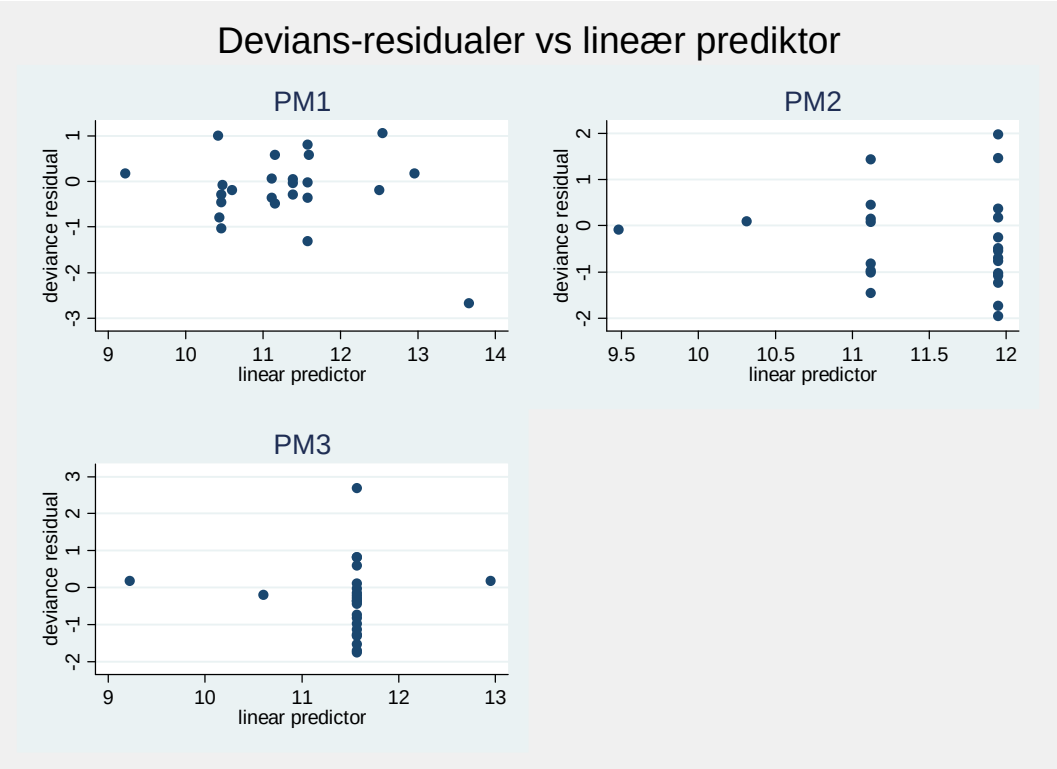
Den totale kovariat-vektoren *U* kan ta 648 forskjellige mulige verdier mens høyst 24 verdier er representert i data (for $X_1 > 0$). Jeg har gjennomført beregningene i (5.1.1) (ikke rapportert her) og fant for enkelte verdier av *U* estimater for $E(X_1 \mid U, Y_1 = 1)$ på opptil 8,5 millioner og enkelte estimater på over 7 millioner – noe som virker urimelig i lys av at maksimalt observert verdi av X_1 er i underkant av 0.7 mill. Siden så få verdier av *U* er representert i data, vil disse resultatene typisk ha karakter av å være ekstrapolasjoner som ofte (i regresjonsanalyse) kan gi misvisende prediksjoner, og som det er vanlig å advare mot i litteraturen. Det at PM1 innholder så mange forklaringsvariable framstår dermed som en svakhet i dette tilfellet (færre dikotome forklaringsvariable gir færre mulige prediksjoner).

Prediksjonsmodellene PM2 og PM3 inneholder ikke *Z*, som innebærer at

$$E(X_1 \mid U, Y_1 = 1) = E(X_1 \mid U, Y_1 = 1, Z = z) \text{ for både } z = 0 \text{ og } 1$$

Kontroll for screening på trinn1 er derfor ikke nødvendig for PM2 og PM3. For PM2 er $U = (ENK, R)$, og for PM3 er $U = (ENK, A_0)$.

Figur 5.1



I figur 5.1 har jeg plottet devians-residualene fra de tre prediksjonsmodellene. I alle tre er den største residualen (i absoluttverdi) eller de to største knyttet til en eller begge av min(X_1) = 8 898 og maks(X_1) = 685 787. I PM1 er den største knyttet til min(X_1), i PM2 er de to største residualene knyttet til hver sin av min(X_1) og maks(X_1). I PM3 er den største knyttet til maks(X_1). Dette er spesielt viktig i det tilfellet at maks og/eller min blant X_1-ene har hatt stor innflytelse på estimeringen, noe som kan føre til redusert troverdighet av predikerte verdier, spesielt ved ekstrapolasjoner. Vi trenger derfor indikasjoner på grad av innflytelse på regresjonsestimeringen for de enkelte observasjonene, for eksempel som i tabell 5.3 der Cook's mål for innflytelse er beregnet for hver observasjon[1].

[1] Cook's mål på innflytelse for observasjon nr. *i* er en veiet sum av endringene i regresjonsestimatene når observasjon *i* fjernes fra data. Målet kan også tolkes som en avstand mellom de to vektorene (med og uten observasjon *i*) av regresjonsestimater.

Tabell 5.3 Cook's mål på innflytelse. Tilfeller av sterk innflytelse er uthevet.

		Cook's mål		
Nr.	X_1	PM1	PM2	PM3
1	8 898	**0.3430**	0.0174	0.0078
2	10 000	0.0358	0.0270	0.0076
3	12 000	0.0502	0.0049	0.0392
4	13 636	0.1163	0.0163	0.0071
5	20 000	0.0326	0.0184	0.0061
6	21 120	0.0111	0.0176	0.0060
7	25 886	0.0048	0.0142	0.0053
8	32 687	0.0019	0.0122	0.0044
9	32 754	0.0502	0.0049	0.0392
10	40 903	0.0288	0.0106	0.0035
11	45 000	0.0179	0.0098	0.0031
12	64 702	0.0064	0.0066	0.0014
13	70 159	0.0004	0.0058	0.0011
14	72 700	0.0050	0.0002	0.0009
15	78 377	**0.2687**	0.0009	0.0006
16	84 300	0.0002	0.0041	0.0004
17	91 033	0.0001	0.0033	0.0002
18	103 200	0.0001	0.0103	0.0000
19	118 140	0.0802	0.0011	0.0001
20	182 195	0.1169	0.0006	0.0048
21	215 855	0.0514	**0.1782**	0.0101
22	217 600	0.0037	0.0032	0.0104
23	500 000	0.0502	0.0970	0.0392
24	685 787	**0.2518**	**0.2296**	**0.2794**

Vi ser at maks(X_1) har sterk innflytelse i alle tre modellene. I PM1 har både min(X_1) og maks(X_1) sterk innflytelse i tillegg til en mellomliggende observasjon (nr. 15). Innflytelsen til min(X_1) er tydelig redusert i PM2 og PM3, som gir økt styrke i forhold til PM1 som prediksjonsmodeller.

Det at den største observerte verdien av X_1 har så stor innflytelse kan også tas som evidens for at gammafordelingen, som ligger til grunn for estimeringen, ikke helt klarer å beskrive tilfredsstillende tendensen til at ekstreme observasjoner dukker opp av og til blant endringsbeløpene i tillegg til graden av ekstremitet (jfr. også figur 5.2 i avsnitt 5.2.1). Standardavvikene bak p-verdiene i tabell 5.1 og 5.2 er de vanlige beregnet ut fra (den observerte) informasjonsmatrisen basert på forutsetningen om gammafordelte endringstall. Det at denne forutsetningen kanskje er tvilsom på grunn av tendensen til ekstreme observasjoner, tilsier at man i stedet burde brukt såkalte "robuste" standardavvik (se siste avsnitt før tabell 5.1). P-verdier basert på robuste standardavvik (kalt "robuste p-verdier") er angitt i parentes i tabell 5.1 og 5.2.

Vi ser også fra figur 5.1 at ikke alle verdier av prediktorene er like godt representert i data. 22 av de 24 virksomhetene er enkeltmannsvirksomheter (ENK) hvorav 21 har ingen ansatte ($A_0 = 1$) som forklarer hvorfor 21 observasjoner har samme verdi for prediktoren i PM3. Dette betyr at hovedtyngden av informasjonen om endringsbeløpene stammer fra

enkeltmannsbedrifter uten ansatte, noe som ytterligere reduserer prediksjonsmulighetene for andre kategorier (se også avsnitt 5.1.2).

I henhold til tabellene 5.1-2 har PM2 og PM3 betydelig bedre prediksjonsegenskaper enn PM1 målt med informasjonskriteriet BIC. Kriteriet AIC skiller ikke vesentlig mellom de tre modellene.

Modellene PM2 og PM3 ble funnet ved å ta utgangspunkt i samme fulle modell som i tabell 5.1, men uten *Z*. Det er derfor naturlig å spørre om *Z* eventuelt kan ha betydning utover de kovariatene i PM2 og PM3. I tabell 5.4 har jeg oppsummert resultatene fra alle 16 submodeller der variablene *ENK, R, A0* og *Z* inngår. En modell i tabellen kan identifiseres ved at den omfatter de kovariatene som har oppgitt p-verdi.

Tabell 5.4 Oppsummering av alle submodeller med kovariater som inngår i PM2 og PM3 (tabell 5.2), med og uten screeningsvariabelen *Z*. Informasjonsmål og robuste p-verdier.

Modell	ENK	R	A0	Z	AIC	BIC
1	0.000	0.152	0.000	0.252	25.28	-38.35
2	0.000	0.073	----	0.756	25.34	-38.11
3	0.000	----	0.000	0.445	25.28	-39.32
4	----	0.103	0.138	0.302	25.38	-36.96
5	0.000	----	----	0.752	25.40	-37.81
6	----	0.070	----	0.677	25.38	-38.19
7	----	----	0.086	0.557	25.43	-36.91
8	----	----	----	0.888	25.46	-37.44
9	0.000	0.248	0.016	----	25.25	-40.27
PM2	0.000	0.058	----	----	25.26	-41.17
PM3	0.000	----	0.000	----	25.23	-41.87
12	----	0.164	0.318	----	25.34	-39.08
13	0.000	----	----	----	25.32	-40.83
14	----	0.078	----	----	25.31	-41.13
15	----	----	0.017	----	25.37	-39.69
16	----	----	----	----	25.38	-40.59

Vi ser at i alle tilfeller når *Z* blir lagt til, vil informasjonskriteriene øke, som betyr at prediksjonsegenskapene, målt med AIC og BIC, reduseres. Heller ikke noen av p-verdiene for *Z* ligger i nærheten av signifikans. Det er således ikke evidens for at *Z* betyr noe vesentlig for endringsbeløpet om ikke ytterligere flere variable legges til slik at det blir overtilpasning, som for eksempel i PM1.

5.1.2 Noen prediksjoner for "endret nettoinntekt", X_1, basert på prediksjonsmodell 2 og 3 fra tabell 5.2

Tabell 5.5 viser estimater for $E(X_1 \mid U, Y_1 = 1)$ for prediksjonsmodell 2 i tabell 5.2.

Tabell 5.5 **Forventet endringsbeløp ("endret nettoinntekt") gitt endring, basert på prediksjonsmodell 2 fra tabell 5.2. Konfidensgrenser basert på robuste standardavvik. Bootstrap (BCa) konfidensgrenser i parentes (4000 replikasjoner). Antall bak observerte gjennomsnitt i parentes.**

					95% konfidensintervall	
Kategori	ENK	Ekstern regnskaps-fører	Observert gjennomsnittlig endringsbeløp gitt endring (1000 kr.)	Forventet endringsbeløp gitt endring (1000 kr)	Nedre konf.grense	Øvre konf.grense
1	Ja	Ja	154 (14)	155	81 (81)	295 (309)
2		Nei	68 (8)	68	36 (35)	127 (133)
3	Nei	Ja	33 (1)	30	20 (--)	45 (--)
4		Nei	12 (1)	13	8 (--)	21 (--)
5	Kat. 3 og 4 slått sammen		22 (2)	22	12 (--)	43 (--)

Resultatene i kategori 5 er oppnådd ved å kjøre glm-regresjon av X_1 med hensyn på *ENK* og samspillsvariabelen $ENK \times R$.

Vi ser at de estimerte forventete endringsbeløpene ligger nær opp til de observerte gjennomsnittene. Dette er et uttrykk for at nesten alle observasjonene stammer fra ENK-virksomheter. Hvis alle hadde vært ENK, ville estimatene vært helt lik de observerte gjennomsnittene. Dette innebærer at estimatene for kategori 3 og 4, som har en observasjon hver, er vesentlig mindre troverdige enn de fra kategori 1 og 2 som har 14 og 8 observasjoner henholdsvis. Det blir som å sammenligne informasjonen i et gjennomsnitt basert på 1 observasjon med et gjennomsnitt basert på 8 eller 14 observasjoner trukket fra populasjoner med omtrent samme varians.

Heller ikke de tilhørende konfidensintervallene i kategori 3 og 4 bør det legges for mye vekt på. De blir for avhengige av den ene observasjonen. Hadde vi hadde vært sikker på at PM2 var sann inklusivt forutsetningen om gammafordeling, ville vi, med en viss troverdighet, kunne brukt de ordinære konfidensintervallene beregnet ut fra informasjonsmatrisen (disse ville blitt (6, 161) for kategori 3, og (2, 72) for kategori 4). Siden vi har grunn til å tvile på at gammafordelingen er den "riktige" fordelingen (jfr. avsnitt 5.1.1), blir det som å beregne et konfidensintervall for forventningen i en fordeling basert på bare en observasjon når vi ikke vet noe om forventningen og variansen i fordelingen.

Usikkerheten med hensyn til forutsetningen om gammafordeling forårsaket bruken av robuste standardavvik for beregning av p-verdier og konfidensintervall. Men vi har et problem til,

nemlig at det her opereres med temmelig små datasett samtidig som at p-verdiene og konfidensintervallene bygger på asymptotisk teori og er først og fremst velbegrunnet for større datasett. Som en kontroll på konfidensgrensene i kategori 1 og 2 beregnet jeg derfor BCa[2] bootstrap konfidensgrenser (basert på 4000 iterasjoner), angitt i parentes i tabellen. Bootstrap intervallene ses å ligge rimelig nær de robuste intervallene, særlig når det gjelder den nedre grensen. De øvre robuste konfidensgrensene virker litt for små men ikke avskrekkende.

For øvrig merker vi oss at det synes å være evidens (jfr. robust p-verdi for *R* i PM2 i tabell 5.2) for at, gitt at det er endring, så er det en tendens til at endringsbeløpet er større blant virksomheter med ekstern regnskapsfører enn blant virksomheter uten. Siden mesteparten av informasjonen om dette stammer fra enkeltmannsvirksomheter, bør man nok være forsiktig med uten videre å overføre konklusjonen til andre virksomhetstyper. Det er også grunn til å minne om forbeholdet som alltid gjelder når man tolker evidens ut fra databestemte prediksjonsmodeller (jfr. forbehold 2 i avsnitt 1 og diskusjonen foran tabell 6.2 i HR): Det er noe evidens for konklusjonen, men ikke så sterk som den hadde vært om modellen hadde vært formulert a priori uavhengig av data og eventuelt i tillegg kvalitetskontrollert via diagnostisk sjekking.

Når det gjelder prediksjonsmodell 3 (PM3) fra tabell 5.2, er tilsvarende resultater vist i tabell 5.6.

Tabell 5.6 **Forventet endringsbeløp ("endret nettoinntekt") gitt endring, basert på prediksjonsmodell 3 fra tabell 5.2. Konfidensgrenser basert på robuste standardavvik. Bootstrap (BCa) konfidensgrenser i parentes (4000 replikasjoner). Antall bak observerte gjennomsnitt i parentes.**

					95% konfidensintervall	
Kategori	ENK	Ingen ansatte	Observert gjennomsnittlig endringsbeløp gitt endring (1000 kr.)	Estimert forventet endringsbeløp gitt endring (1000 kr)	Nedre konf.grense	Øvre konf.grense
1	Ja	Ja	105 (21)	106	59 (65)	191 (218)
2		Nei	500 (1)	421	305 (--)	583 (--)
3	Nei	Ja	12 (1)	10	7 (--)	14 (--)
4		Nei	33 (1)	40	27 (--)	61 (--)
5	Kat. 2, 3 og 4 slått sammen		272 (3)	182	43 (--)	762 (--)

Kategori 5 er oppnådd ved å kjøre glm-regresjon av X_1 med hensyn på samspillsvariabelen $ENK \times A_0$.

Trekkene i tabell 5.6 ligner på dem som er beskrevet for foregående tabell 5.5. De estimerte forventete endringsbeløpene ligger også her nær de observerte gjennomsnittene. Dette gjør

[2] BCa står for "bias corrected and accelerated", og viser til en type bootstrap konfidensintervall som ofte anbefales.

resultatene for kategori 2, 3 og 4, med bare en observasjon hver, mindre troverdige (spesielt tydelig for kategori 2) i og med at modellen ikke klarer å fange opp variasjonen i disse gruppene. De tre kategoriene er derfor slått sammen til en kategori 5. Det er interessant at konfidensintervallet blir stort for denne kategorien, noe som virker rimelig.

Bootstrap konfidensintervall for kategori 2 – 5 mangler på grunn av for få observasjoner i hver kategori. For kategori 1, enkeltmannsvirksomheter uten ansatte, er intervallet forskjøvet litt mot høyre i forhold til intervallet basert på robuste standardavvik.

P-verdiene i tabell 5.2 for PM3 kan tyde på en tendens til høyere endringsbeløp i virksomheter med en eller flere ansatte enn i virksomheter uten ansatte. Dette synes å være i tråd med den tilsvarende konklusjonen om ekstern regnskapsfører etter tabell 5.5.

Det er naturlig å spørre om man kan kombinere informasjonen i tabell 5.5 og 5.6. For eksempel, hva vil det predikerte gjennomsnittbeløpet være for en enkeltmannvirksomhet uten ansatte og uten ekstern regnskapsfører, gitt at det er endring? Et rimelig svar ville være det samme som i tabell 5.6 for en enkeltmannsvirksomhet uten ansatte, nemlig 106 tusen. Modell 9 i tabell 5.4 viser at data ikke gir noe evidens for at *R* betyr noe utover *ENK* og A_0, så prediksjonen bør være den samme enten virksomheten har ekstern regnskapsfører eller ikke.

Vi ser av tabell 5.6 at hovedtyngden av informasjonen i data stammer fra enkeltmannsvirksomheter uten ansatte (med 21 observasjoner av i alt 24). Hvis vi derfor definerer delpopulasjonen

"*Gruppe A*" som *enkeltmanns-virksomheter uten ansatte,*

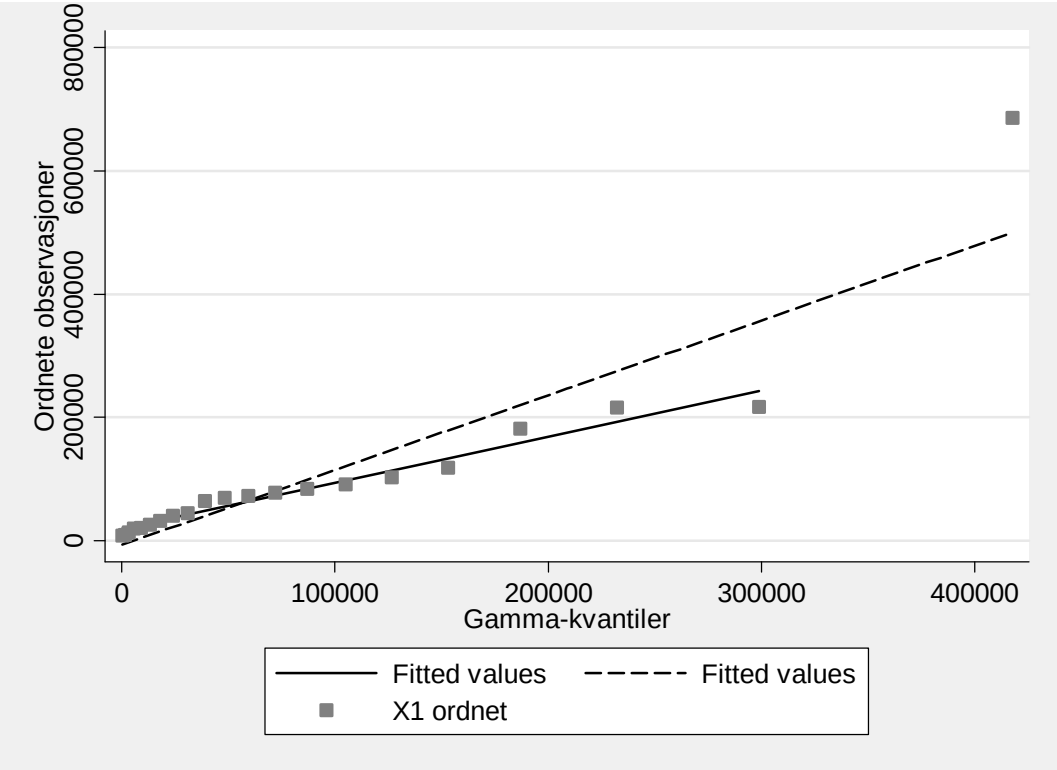
vil bare tre av de 24 endringstilfellene av typen "endret nettoinntekt" stamme fra andre kategorier enn gruppe A. Tabell 5.6 viser også at prediksjoner utenfor denne gruppen av gjennomsnittlig endringsbeløp gitt endring har liten troverdighet. Dette burde tilsi en ny analyse for denne gruppen isolert helt fra grunnen av, med analyse av *Z*, estimering av endringssannsynligheter og forventet endringsbeløp. På grunn av tidsbegrensninger er dette ikke gjennomført i denne rapporten, med unntak av litt om endringsbeløp nedenfor, men kan eventuelt bli gjenstand for en senere rapport.

Jeg gjennomførte en tilsvarende analyse som i avsnitt 5.1 for de 21 observasjonene i gruppe A, med å starte med full modell og søke etter utsagnskraftig prediksjonsmodeller, men fant ingen evidens (ikke rapportert her) for at fordelingen for X_1 gitt endring i gruppe A, avhenger av noen av de foreliggende kovariatene[3]. Det vil si at den "beste" prediksjonsmodellen er den som forutsetter konstant forventning, dvs. med konstant lineær prediktor, og estimatet blir identisk med estimatet i tabell 5.6 (dvs. 106 tusen) og med samme konfidensgrenser som i tabell 5.6. Hvis vi antar at observasjoner av X_1 er uavhengige og identisk fordelte i gruppe A, som det således er evidens for, kan vi se mer eksplisitt på forutsetningen om gammafordeling som vi fant grunn til å tvile på ovenfor. I figur 5.2 har jeg laget et kvantilplott med de observerte X_1-ene ordnet etter størrelsen på Y-aksen og gamma-kvantiler basert på den estimerte gammafordelingen på X-aksen. Hvis gammafordelingen er realistisk, vil

[3] Det dukket også i dette tilfellet opp en klart overtilpasset prediksjonsmodell med mange (7) "signifikante" kovariater og ekstremt høy (0.995) modell p-verdi. Denne prediksjonsmodellen ble også forkastet på grunn av overtilpasning.

observasjonene føye seg tilnærmet til en rett linje. Jeg har også tegnet inn regresjonslinjer med og uten den ekstreme observasjonen. Den heltrukne linjen viser at observasjonene uten

Figur 5.2 Kvantilplott for X_1 (endret nettoinntekt) mot gammafordelingen i gruppe A



den ekstreme observasjonen (686 tusen) synes å føye seg godt til en gammafordeling. Den stiplete linjen er beregnet der den ekstreme observasjonen er tatt med, og vi ser da at de sentrale observasjonene synes å bli for små for en gammafordeling og/eller at den ekstreme blir for stor i forhold til en gamma-fordeling. Vi har således fått en ny bekreftelse på at gammafordelingen ikke i tilstrekkelig grad klarer å fange opp størrelsen på ekstreme observasjoner som har en tendens til å dukke opp en gang i blant.

Noen avsluttende merknader om PM1 i tabell 5.1: Som nevnt synes PM1 uegnet siden den kan gi mange lite troverdige prediksjoner. For eksempel, for en enkeltmannsvirksomhet med ekstern regnskapsfører og flere ansatte og omsetning under 1 mill, og som kommer fra en hovedsakelig tjenesteytende kommune utenfor Skatt øst og vest, så blir predikert endringsbeløp X_1 gitt endring, 8.5 mill i henhold til ligning 5.1.1. I følge tabell 5.6 imidlertid vil virksomheten høre med i kategori 5, og prediksjonen 182 tusen med betydelig usikkerhet (konfidensintervall (43, 762)), er et mer troverdig resultat. Siden, imidlertid, mesteparten av informasjonen om endringsbeløpene stammer fra enkeltmannsvirksomheter uten ansatte (kategori 1), kunne man argumentere at PM1 kun burde brukes innenfor kategori 1. Innenfor denne kategorien blir de predikerte gjennomsnittlige endringsbeløpene noe rimeligere med minste prediksjon 2.9 tusen og maks prediksjon 442 tusen. Men fortsatt gjelder en utstrakt grad av ekstrapolasjon ved disse prediksjonene siden kategori 1 omfatter 162 mulige kombinasjoner av verdier for kovariatene i ligning 5.1.1 mens bare høyst 21 er representert i data.

Diskusjonen for X_2 i avsnitt **5.2.1** viser at PM1 kan karakteriseres som et eksempel på overtilpasning ("over fitting"), som er velkjent fra vanlig regresjonsanalyse og som bør unngås. Et ekstremt tilfelle i vanlig regresjonsanalyse er når vi bruker like mange forklaringsvariable (inklusivt konstantleddet) som antall observasjoner. Da får vi perfekt tilpasning av modellen - men en modell med svært dårlige prediksjonsegenskaper. Et symptom på at vi er i en tilsvarende situasjon her med PM1, er den høye p-verdien (0.9998) for testing av redusert modell mot full modell (jfr. tabell 5.1).

5.2 Separat analyse av X_2 (beløp for "endret merverdiavgift")

Datagrunnlaget for X_2 er kun 13 observasjoner som er litt tynt for en omfattende regresjonsanalyse. Blant de 13 er det for eksempel ingen "fotografer" og kun en "jernvarehandel", så analysen vil først og fremst ha relevans for bransjene "snekker" og "design". Om man ønsker å overføre resultater fra analysen til bransjene "fotograf" og "jernvare", vil man måtte bygge på en antakelse at det ikke er forskjell mellom bransjene – en antakelse det ikke er informasjon i data til å teste. Tilsvarende gjelder for region i og med at det er null observasjoner fra Skatt nord og bare en fra Skatt vest.

I tillegg fins det en ekstrem observasjon, 612 tusen, mens maksimum av de øvrige ligger på 40 tusen.

En negativ observasjon antas utypisk for interesse-populasjonen og fjernes slik at det blir igjen 13 observasjoner av X_2.

Med dette litt begrensete utgangspunktet har jeg likevel forsøkt å gjennomføre analysen som før - med full modell og valg av prediksjonsmodeller basert på statistiske kriterier ut fra data. Resultatet er gitt i tabell 5.7 som viser to alternative prediksjonsmodeller, PM1 og PM2. Mitt endelige forslag til prediksjonsmodell er PM2, da PM1 viser seg å bære preg av såkalt *overtilpasning*, noe som er en klar svakhet og innebærer reduserte prediksjonsegenskaper. En nærmere diskusjon av dette er gitt i avsnitt 5.2.1. Grunnen til at jeg likevel har referert den her er de små p-verdiene for regresjonskoeffisientene som kan virke litt overraskende (sammenlignet med typiske symptomer på overtilpasning i vanlig regresjonsanalyse), og som kan tjene som en advarsel om at det ikke alltid er nok å se på p-verdiene for regresjonskoeffisentene alene ved vurdering av egnetheten til en prediksjonsmodell. Symptomet på overtilpasning i dette tilfellet er først og fremst den høye p-verdien, 0.992, for likelihood- ratio-testen (LR) av prediksjonsmodellen mot den fulle modellen (selv om LR-testen ikke er helt velbegrunnet her som test betraktet på grunn av usikkerheten knyttet til gammafordelingen). Informasjonskriteriene AIC og BIC indikerer også bedre prediksjonsegenskaper for PM2 enn PM1.

Tabell 5.7 **Regresjonsresultater (GLM, gamma og log link) for endringsbeløp av typen "endret merverdiavgift" gitt endring og funn/ikke-funn på trinn 1. Robuste p-verdier.**

(Basert på utskrift Ut19, Ut20 og Ut21 i appendiks A7)

Avhengig X_2	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
AS	----	----	----	----	----	----
ENK	-2.2985	0.000	-1.3115	0.001	-2.7360	0.000
Ost	-0.0107	0.974	----	----	----	----
Sor	----	----	----	----	----	----
Vest	----	----	----	----	----	----
Midt	0.9789	0.000	----	----	----	----
Nord	----	----	----	----	----	----
Snekker	-0.4742	0.156	----	----	----	----
Jernv	-2.9725	0.003	-1.9812	0.000	----	----
Fotograf	----	----	----	----	----	----
Design	----	----	----	----	----	----
Z	0.1501	0.654	----	----	----	----
R	0.0767	0.819	----	----	----	----
Nyreg	2.4251	0.000	1.9112	0.000	2.0013	0.000
Oms0_10	-2.4323	0.000	-2.2707	0.000	-1.3825	0.000
A0	----	----	----	----	----	----
A1	----	----	----	----	----	----
Komtjenest	-0.4270	0.202	-1.0942	0.000	----	----
Komsentral	1.5127	0.000	1.8218	0.000	----	----
konstant	12.0696	0.000	11.5023	0.000	12.6761	0.000
Antall obs	13		13		13	
Log pseudo-likelihood	-138.081		-138.327		-140.636	
Devians	3.4994		3.9927		8.6098	
-2 log LR			0.4932		5.1104	
P-verdi redusert vs full modell			0.992		0.746	
AIC	23.09		22.36		22.25	
BIC	0.93		-11.40		-14.47	

Det er ingen evidens for at screeningen på trinn 1 (*Z*) har noen effekt på fordelingen for X_2. Legger man for eksempel *Z* til kovariatene i PM2, får *Z* en p-verdi på 0.32, mens p-verdiene for de andre er nesten uberørte.

Tabell 5.8 **Forventet endringsbeløp gitt endring for X_2 basert på prediksjonsmodell 2 i tabell 5.7. Konfidensgrenser basert på robuste standardavvik. Prediksjoner for kategorier uten observasjoner (satt i parentes) er tvilsomme. Antall bak observerte gjennomsnitt i parentes.**

						95% Konfidensintervall	
Kategori	ENK	Nyregistrert	Omsetning under 1 mill.	Observert gjennomsnittlig endringsbeløp gitt endring (1000 kr.)	Estimert forventet endringsbeløp gitt endring (1000 kr)	Nedre konf. grense	Øvre konf. grense
1	Ja	Ja	Ja	39 (2)	39	37	40
2			Nei	-- (0)	(153)	(73)	(325)
3		Nei	Ja	5 (6)	5	3	9
4			Nei	21 (3)	21	12	35
5	Nei	Ja	Ja	-- (0)	(594)	(144)	(2447)
6			Nei	-- (0)	(2368)	(572)	(9802)
7		Nei	Ja	-- (0)	(80)	(18)	(365)
8			Nei	320 (2)	320	86	1192

Tabell 5.8 gir en god illustrasjon på hvor lite prediktiv informasjon som er i data. Det virker som om hver kategori ligger isolert i forhold til hverandre med lite felles informasjon slik at hver predikert verdi blir lik det observerte gjennomsnittet i de fire kategoriene der det foreligger observasjoner. Dessuten virker prediksjonene for kategorier uten observasjoner lite troverdige (satt i parentes i tabellen). Siden estimeringen således leder til vanlige gjennomsnitt i hver kategori, betyr det at informasjonsgrunnlaget for prediksjoner i kategorier med bare 2 eller 3 observasjoner er tynt. Litt bedre fatt er det for kategori 3 som har 6 observasjoner og predikert verdi 5 tusen.

På den annen side er antakelig forventet beløp i kategori 3 betydelig underestimert hvis det er slik at tendensen til at ekstreme (store) observasjoner dukker opp en gang i blant, er en generell tendens som gjelder alle kategorier. I så fall kan prediksjonene i kategori 1, 3 og 4, som ikke har ekstreme observasjoner, være betydelig undervurdert, mens prediksjonen i kategori 8 (som inneholder den ekstreme observasjonen på 612 tusen) kan være betydelig overvurdert siden den har for få "vanlige" observasjoner på under 50 tusen. En indikasjon på disse over- og undervurderingene kan vi få ved å anta at alle kategorier har samme forventet beløp. I så fall blir estimatet på forventet beløp lik gjennomsnittet på 62 tusen. Et 95% bootstrap konfidensintervall (BCa) for forventet beløp ble (14 tusen, 245 tusen) basert på samme antakelse. I mangel av en god teori for hvor ofte ekstreme observasjoner dukker opp og deres fordeling, er det foreløpig ikke stort mer vi kan si om prediksjoner i dette tilfellet.

Kan andre tendenser leses ut av data? Om vi kjører glm-regresjon av X_2 med hensyn på konstantleddet alene, får vi en devians på 40.39 som sammenlignet med deviansen til den fulle modellen 3.50 gir en -2 ganger pseudo likelihood ratio på 36.89. Med 11 frihetsgrader gir dette en p-verdi på 0.000 (basert på kji-kvadrat fordelingen med 11 frihetsgrader) for hypotesen at forventet beløp er konstant i alle kategorier. Hvis forutsetningen om gammafordeling hadde vært realistisk ville dette implisere at det er evidens i data for at det er avhengighet tilstede mellom forventete beløp og kovariatene. På grunn av usikkerheten om modellen med gammafordeling er ikke denne testen helt velbegrunnet, men resultatet kan likevel tas som en viss evidens for en slik avhengighet og at det dermed kan være noe hold i

signifikansene i tabell 5.7 for prediksjonsmodell 2. (P-verdiene for koeffisientene i prediksjonsmodell 1 er mer tvilsomme på grunn av overtilpasning). I så fall er det grunnlag for å si at det er noe evidens i data for at forventet endringsbeløp gitt endring er høyere for ikke-ENK virksomheter (hovedsakelig AS) enn for ENK-virksomheter, at den er høyere for nyregistrerte (inntil 3 år) enn for eldre virksomheter, samt at den er høyere for virksomheter med omsetning over en million enn om omsetningen er under en million.

5.2.1 Mer om overtilpasning for prediksjonsmodell 1 fra tabell 5.7

Tabell 5.9 viser alle observasjonene for X_2 sammen med tilhørende estimater for forventet verdi samt kovariater for prediksjonsmodell 1 i tabell 5.7. I tillegg vises standardavvikene for den linære prediktoren (dvs. for log(forventet verdi)). Observasjonene er ordnet etter størrelsen på estimert forventning. Observasjoner som har samme estimert forventning har også (i dette tilfellet) samme kombinasjon av kovariat-verdier.

Tabell 5.9 **Sammenligning av observert verdi av X_2 og estimert forventet verdi for prediksjonsmodell 1 i tabell 5.7. Ordnet etter forventet verdi.**

				Kovariater i prediksjonsmodell 1 (tabell 5.7)					
Nr.	Observert X_2	Estimert Forventet X_2	St.avvik Lineær prediktor	Jernv	ENK	Nyreg	Oms0_10	Komtje~t	Komsen~l
1	2 752	2 751.999	0.0000	0	1	0	1	0	0
2	8 771	5 697	0.2829	0	1	0	1	1	1
3	10 000	5 697	0.2829	0	1	0	1	1	1
4	3 520	5 697	0.2829	0	1	0	1	1	1
5	5 717	5 697	0.2829	0	1	0	1	1	1
6	477	5 697	0.2829	0	1	0	1	1	1
7	8 925	8 925	-----	0	1	0	0	1	0
8	31 333	26 657	0.1291	0	1	0	0	0	0
9	21 980	26 657	0.1291	0	1	0	0	0	0
10	28 246	28 246	-----	1	0	0	0	1	1
11	39 787	38 518	0.0242	0	1	1	1	1	1
12	37 249	38 518	0.0242	0	1	1	1	1	1
13	611 774	611 774	-----	0	0	0	0	0	1

Vi ser at tre observasjoner (nr. 7, 10 og 13) har perfekt tilpasning (med samme observert og estimert verdi) og en observasjon (nr. 1) nesten perfekt tilpasning. Vi har altså perfekt eller nesten perfekt tilpasning i alle de fire tilfellene der det bare er en observasjon fra kategorien bestemt av kovariatene. Hvis vi hadde hatt like mange variable (inklusivt konstantleddet) som antall observasjoner, ville vi fått perfekt tilpasning for alle observasjonene som er et ekstremt tilfelle av overtilpasning. Selv om modellen i så fall beskriver observasjonene perfekt, vil en prediksjon fra en slik modell være nesten verdiløs. En enkelt observasjon trukket fra en populasjon er ikke nok til å si noe om variasjonen i populasjonen så sant det ikke er en link mellom forventning og varians. Dette kommer til uttrykk i tabellen med ubestemte standardavvik der det er perfekt tilpasning. En slik modell (med for mange

forklaringsvariable) har derfor dårligere prediksjonsegenskaper enn den høye tilpasningsgraden skulle tilsi – som er bakgrunnen for uttrykket "overtilpasning".

Et symptom på overtilpasning kan være en høy p-verdi (nær 1) for testing av prediksjonsmodellen mot den fulle modellen. Vi ser dessuten fra tabell 5.9 at estimerte varianser for perfekt eller nesten perfekt tilpassede prediksjoner er nær null eller ikke-eksisterende. Dette kan føre til en betydelig undervurdering av kovariansmatrisen for regresjonsestimatorene, som kan være en forklaring på at vi får så påfallende mange kovariater med lave p-verdier for prediksjonsmodell 1 ut fra bare 13 observasjoner – noe som dermed også vil kunne tolkes som et symptom på overtilpasning.

I en slik situasjon med overtilpasning bør man således være forsiktig med å legge for mye vekt på variable med signifikante p-verdier siden bruken av dem kan lede til sterkt misvisende prediksjoner for verdikombinasjoner av kovariatene som forekommer sjelden eller ikke i det hele tatt i data.

5.3 Separat analyse av X_3 (beløp for "påvist uteholdt omsetning")

Datagrunnlaget for "*påvist uteholdt omsetning*" (X_3) er 16 observasjoner. Alle disse er enkeltmannsvirksomheter (bortsett fra en VIFE), alle uten ansatte som jeg kalte "gruppe A" i avsnitt 5.1.2. Om vi utvider gruppe A litt og definerer

"*Gruppe B*" - *Ikke-AS virksomheter uten ansatte*,

har vi altså kun observasjoner av X_3 fra gruppe B hvorav alle bortsett fra en, er fra gruppe A. Erfaringene fra avsnitt 5.1 og 5.2 tilsier at vi bør være forsiktige med å predikere endringsbeløp for grupper uten observasjoner i dette materialet.

Tabell 5.10 **Regresjonsresultater (GLM, gamma og log link) i gruppe B for endringsbeløp av typen "påvist uteholdt omsetning" gitt endring og funn/ikke-funn på trinn 1. Robuste p-verdier.**

(Basert på utskrift Ut22 og Ut23 i appendiks A8)

Avhengig X_3	Full modell		Prediksjonsmodell 1	
	Koeff.	p-verdi	Koeff.	p-verdi
AS	-----	-----	-----	-----
ENK	-----	-----	-----	-----
Ost	-0.7216	0.005	-1.4203	0.000
Sor	2.0656	0.801	-----	-----
Vest	-----	-----	-----	-----
Midt	-1.0914	0.565	-----	-----
Nord	-----	-----	-----	-----
Snekker	4.0006	0.545	-----	-----
Jernv	-2.6855	0.745	-2.4753	0.000
Fotograf	1.8049	0.534	-----	-----
Design	-----	-----	-----	-----

Z	-1.4677	0.829	-----	-----
R	-3.3975	0.801	-----	-----
Nyreg	1.8069	0.792	-----	-----
Oms0_3	0.3464	0.961	-----	-----
Oms3_10	-0.5275	0.938	-----	-----
Oms0_10	-----	-----	-----	-----
A0	-----	-----	-----	-----
A1	-----	-----	-----	-----
Komtjenest	2.5930	0.711	-----	-----
Komsentral	-3.0269	0.653	-----	-----
konstant	11.7005	0.396	11.3878	0.000
Antall obs	16		16	
Log-pseudo likelihood	-183.1286		-186.1525	
-2 log LR			6.0477	
P-verdi redusert vs full modell			0.870	
AIC	24.64		23.64	
BIC	4.68		-19.78	

Hullene i den fulle modellen skyldes kollinearitetsproblemer samt at alle observasjonene stammer fra Gruppe B.

Jeg fant ingen evidens i data for at screeningsvariabelen *Z* har betydning. *Z* oppnådde ikke signifikans i noen av de reduserte modellene jeg så på og tilstedeværelse av *Z* førte i alle tilfellene til at informasjonskriteriene AIC og BIC økte. Vi vil således anta at

$$E(X_3 \mid U, Z, Y_3 = 1) = E(X_3 \mid U, Y_3 = 1) \text{ der } U = (Ost, Jernv).$$

Det er altså ikke grunn til å kontrollere for *Z* i dette tilfellet.

Prediksjoner basert på prediksjonsmodellen i tabell 5.10 er vist i tabell 5.11.

Tabell 5.11 **Forventet endringsbeløp ("påvist uteholdt omsetning") i gruppe B, gitt endring, basert på prediksjonsmodellen i tabell 5.10. Konfidensgrenser basert på robuste standardavvik. Bootstrap (BCa) konfidensgrenser i parentes (4000 replikasjoner). Antall i parentes.**

					95% konfidensintervall	
Kategori	Skatt Øst	Bransje "Jernvare-handel"	Observert gjennomsnittlig endringsbeløp gitt endring (1000 kr.)	Estimert forventet endringsbeløp gitt endring (1000 kr)	Nedre konf.grense	Øvre konf.grense
1	Ja	Ja	-- (0)	2	1 (--)	5 (--)
2		Nei	21 (5)	21	15 (--)	30 (--)
3	Nei	Ja	7 (2)	7	3 (--)	18 (--)
4		Nei	88 (9)	88	49 (43)	160 (145)
5	Kat. 2 og 4 slått sammen		64 (14)	64	36 (33)	116 (111)
6	Alle kategorier slått sammen		57 (16)	57	31 (32)	105 (103)

De tre største observasjonene (rundt 200 tusen) ligger alle i kategori 4 mens de øvrige ligger godt under 100 tusen. Dermed, siden datamaterialet er så lite, er det også her sannsynlig at estimatet i forventet endringsbeløp er overestimert i kategori 4 og underestimert i kategori 1, 2 og 3 (jfr. diskusjonen etter tabell 5.8). Antar vi at alle kategorier har samme forventet endringsbeløp gitt endring (kategori 6), blir estimatet på forventet beløp gitt endring for gruppe B lik gjennomsnittet på 57 tusen. Et 95% bootstrap konfidensintervall (BCa) for forventet beløp ble (32 tusen, 103 tusen) basert på samme antakelse. I mangel av en god teori for hvor ofte ekstreme observasjoner dukker opp og deres fordeling, er dette antakelig den beste prediksjonen vi kan gjøre for gruppe B (eventuelt gruppe A) i dette tilfellet.

For øvrig kan det synes å være en viss (svak) evidens for at forventet beløp gitt endring i gruppe B er lavere i Skatt Øst enn i andre regioner. Dette gjelder hvis vi skal tro på de lave p-verdiene for regresjonskoeffisienten i tabell 5.10. Den høye p-verdien (0.87) for LR-testen for prediksjonsmodellen samt det at de observerte og predikerte gjennomsnittene i tabell 5.11 er like, kan tyde på en viss grad av overtilpasning - noe som kan ha påvirket p-verdiene for regresjonskoeffiesientene via en underestimert kovariansmatrise for koeffisient-estimatorene. Tilsvarende konklusjon for "*Jernvarehandel*" – dvs. at denne bransjen har en tendens til lavere verdier enn i andre bransjer – virker mindre troverdig, til tross for den lave p-verdien i tabell 5.10, siden "*Jernvarehandel*" bare er representert med 2 observasjoner.

5.4 Separat analyse av X_4 (beløp for "endret nettoinntekt" av andre typer enn "påvist uteholdt omsetning")

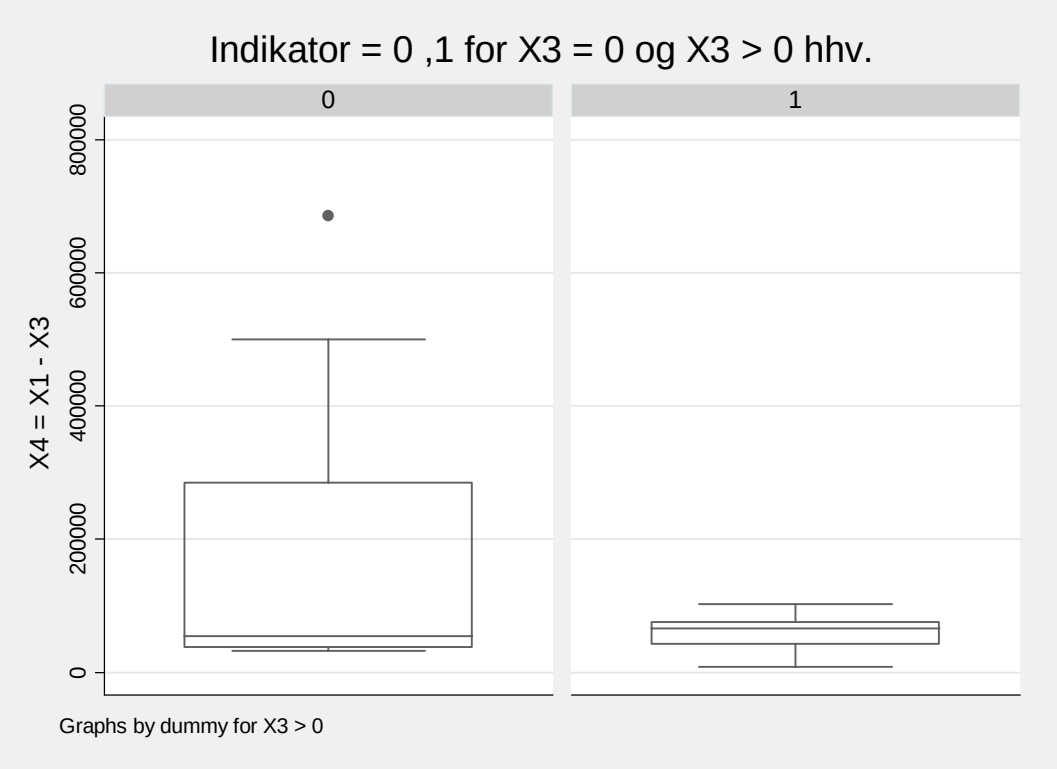
I avsnitt 4.8 diskuterte vi sannsynligheten for "*endret nettoinntekt*" av andre typer enn "*påvist uteholdt omsetning*", som inntreffer når $X_1 > X_3$, og måles med $X_4 = X_1 - X_3$.
X_4 har 14 observasjoner større enn 0, hvorav 12 er i gruppe A (enkeltmannsvirksomheter uten ansatte). Tabell 5.12 viser en oversikt.

Tabell 5.12 Oversikt over X_4 gitt endring i diverse subgrupper.

X_4	Antall > 0	Gj.snitt	St.avvik	Median	Min	Maks
Gruppe A	12	108 214	183 496	61 068	8 220	685 787
Ikke gruppe A	2	266 377	330 393	330 393	32 754	500 000
$X_3 = 0$	8	183 999	257 573	54 851	32 687	685 787
$X_3 > 0$	6	59 889	32 265	65 567	8 220	102 140

I og med at $X_1 = X_3 + X_4$, har vi således splittet opp beløp for "*endret nettoinntekt*" i to typer, beløp for "*påvist uteholdt omsetning*" (X_3) og andre typer (X_4). Det er nå naturlig å spørre om det er rimelig å anta at X_3 og X_4 er stokastisk uavhengige. Det ville i så fall lette oppgaven å estimere en simultan fordeling for (X_3, X_4). Figur 5.3 indikerer imidlertid en betydelig forskjell i den betingete fordelingen til X_4 gitt at $X_3 = 0$ eller gitt at $X_3 > 0$, noe som impliserer avhengighet mellom X_3 og X_4 (jfr. også tabell 5.12).

Figur 5.3 **Boks-plott for $X_4 > 0$ når $X_3 = 0$ (8 obs.) og når $X_3 > 0$ (6 obs.)**



På grunn av det tynne materialet har jeg derfor ikke gjort noe forsøk på å estimere simultanfordelingen for X_3 og X_4.

Når det gjelder marginalfordelingen til X_4, har jeg videre, i lys av det magre datagrunnlaget utenfor gruppe A, kun sett på de 12 observasjonene fra gruppe A.

I gruppe A er det en ekstrem observasjon på 686 tusen. Det dreier seg om en design-virksomhet i Skatt øst som ikke er nyregistrert og med positiv omsetning under 300 tusen. I tillegg, som medlem av gruppe A, er det en enkeltmanns-virksomhet uten ansatte. Som illustrert for de andre beløps-variablene, vil en slik observasjon ha stor innflytelse på estimering og tolkning av resultatene når datasettene er så små som her. En tendens synes å være overestimering av forventet endringsbeløp gitt endring i de grupper som inneholder den ekstreme observasjonen, og underestimering i de grupper som ikke inneholder den ekstreme observasjonen. I lys av dette, samt tendensen for overtilpasning i mange submodeller, vil jeg her ta en mer deskriptiv tilnærming til analysen av X_4.

I tabell 5.13 vises resultatene av å kjøre enkle glm-regresjoner av X_4 med hensyn på kovariatene hver for seg. P-verdien for regresjonskoeffisienten og informasjonskriteriene AIC og BIC er vist for hver regresjon. For hver kovariat vises også hvor mange observasjoner av de 12 som er i gruppen bestemt av verdi 1 for kovariaten. For eksempel, Skatt Øst har 5 observasjoner med gjennomsnitt og maks verdi for X_4 vist i de to siste kolonnene.

Tabell 5.13 **Resultater fra enkle glm-regresjoner (gamma med log-link) for X_4 i gruppe A. Robuste p-verdier.**

		Glm regresjon av X_4 på kovariatene hver for seg				
Kovariat	Antall der kovariat =1	P-verdi for regr.-koeff.	AIC	BIC	Gj.snitt X_4	Maks X_4
Ost	5	0.065	25.15	-15.02	183 790	685 787
Sor	4	0.053	25.25	-13.85	42 765	75 529
Vest	1	0.284	25.50	-10.89	64 702	64 702
Midt	2	0.388	25.50	-10.98	71 930	73 700
Nord	0	---	---	---	---	---
Snekker	7	0.111	25.22	-14.20	59 003	102 140
Jernv	1	0.456	25.51	-10.77	75 529	75 529
Fotograf	2	0.211	25.46	-11.37	58 006	73 700
Design	2	0.017	24.94	-17.63	347 004	685 787
Z	7	0.200	25.37	-12.44	140 821	685 787
Nyreg	1	0.011	25.43	-11.69	32 687	32 687
Ikke Nyreg	11	---	---	---	115 081	685 787
R	9	0.263	25.45	-11.51	123 003	685 787
Ikke R	3	---	---	---	63 847	75 529
Oms03	5	0.188	25.29	-13.36	168 907	685 787
Oms310	3	0.273	25.45	-11.45	65 278	73 700
Oms0_10	8	0.267	25.42	-11.84	130 046	685 787
Komtjenest	7	0.161	25.34	-11.20	142 981	685 787
Komsentral	9	0.202	25.43	-11.70	124 301	685 787

Vi legger merke til at nesten alle (11 av 12) av endringstilfellene (i gruppe A) er *ikke-nyregistrerte* (eksistert i minst 4 regnskapsår), de fleste (9 av 12) stammer fra Skatt Øst eller Sør og er dominert (7 av 12) av bransjen "*snekkerarbeid*". I tillegg, til tross for at de alle er enkeltmannsvirksomheter uten ansatte, har de fleste av dem (9 av 12) ekstern regnskapsfører. Disse tendensene er naturligvis ikke signifikante i et så lite materiale, men gir likevel en pekepinn om hva slags virksomheter som synes å være mest aktuelle.

Hvis vi ser på de enkeltvise regresjonsresultatene, er tilsynelatende regresjonen av X_4 med hensyn på *Design* "vinneren" målt med informasjonsmålene og p-verdi for regresjonskoeffisienten. Det framgår imidlertid at det bare er 2 observasjoner i data fra design-bransjen hvorav den ekstreme observasjonen er den ene. På grunn av størrelsen på den ekstreme observasjonen blir kontrasten mellom design-gruppen og de andre 10 klart overestimert, og bruk av *Design*-dummien som prediktor vil gi misvisende resultater. Det er også interessant å legge merke til at den ekstreme observasjonene i dette tilfellet har hatt en utilbørlig stor innflytelse på informasjonskriteriene, spesielt på BIC.

Når det gjelder regresjonen på *Nyreg* (dummy for nyregistrert), er situasjonen litt annerledes siden den ekstreme observasjonen hører med blant de 11 observasjonene for ikke-nyregistrerte virksomheter og er dermed bedre balansert av de 10 "vanlige" observasjonene. Det betyr at *Nyreg* er aktuell som prediktor – via kategorien ikke-nyregistrerte virksomheter.

I tillegg til *Nyreg* er andre aktuelle prediktorer basert på tabell 5.13, *Skatt Øst og Sør, Snekker, R, Z, oms03, Komtjenest, Komsentral*. Jeg gjennomførte en søkeprosedyre basert på disse (ikke rapportert) på samme måte som før og endte opp med *Nyreg* og *Øst* som det tilsynelatende beste utvalget av prediktorer, med følgende estimerte regresjonsfunksjon (p-verdier i parentes):

$$\text{(PM1)} \qquad \log\left(\hat{E}(X_4 \mid Nyreg, Øst)\right) = \underset{(0.000)}{10.901} \underset{(0.003)}{-1.914 \cdot Nyreg} \underset{(0.032)}{+1.407 \cdot Øst}$$

Jeg fant ingen evidens i data for at screeningsvariabelen, *Z*, hadde betydning utover *Nyreg* og *Øst*, slik at det ikke skulle være nødvendig å kontrollere for denne.

Prediksjonsmodellen PM1 ser i utgangspunktet bra ut, man har likevel svakheter – noe som framgår av tabell 5.14 - over predikerte verdier:

Tabell 5.14 Forventet endringsbeløp gitt endring ("endret nettoinntekt" av andre typer enn "påvist uteholdt omsetning") i gruppe A, basert på prediksjonsmodellen PM1. Konfidensgrenser basert på robuste standardavvik. Antall i parentes.

					95% konfidensintervall	
Kategori	Nyregistrert	Skatt Øst	Observert gjennomsnittlig endringsbeløp gitt endring (1000 kr.)	Estimert forventet endringsbeløp gitt endring (1000 kr)	Nedre konf.grense	Øvre konf.grense
1	Ja	Ja	33 (1)	33	----	----
2		Nei	-- (0)	----	----	----
3	Nei	Ja	222 (4)	222	64	768
4		Nei	54 (7)	54	39	75

Vi ser at de estimerte verdiene er eksakt lik de observerte gjennomsnittene slik at det er ikke noe fordeling/overføring av informasjon mellom kategoriene. Dermed blir prediksjonene nærmest ubrukelige i kategori 1 og 2 (nyregistrerte virksomheter). Også for kategori 3 og 4 er prediksjonene tvilsomme siden den ekstreme observasjonen, som tilhører kategori 3, klart har for stor innflytelse. Siden de forventete endringene er lik de observerte gjennomsnittene, er det således grunn til å tro at forventet endring kan være betydelig overestimert i kategori 3 og betydelig underestimert i kategori 4. Vi kan derfor konkludere at prediksjonsmodell PM1 ikke synes velegnet for prediksjonsformål.

Er det, på den annen side i lys av den lave p-verdien for *Øst* i PM1, rimelig å si at det er evidens for en tendens til at X_4 tar høyere verdier i Skatt øst enn andre steder? Svaret på det vil være nokså uavklart i dette tilfellet. Det er grunn til å tro at den ekstreme observasjonen har bidratt for mye til den lave p-verdien til at vi kan ta den helt alvorlig. Om vi for eksempel sammenligner de fire observasjonene fra kategori 3 med de sju fra kategori 4 ved hjelp av en Wilcoxon test eller en median test, som ikke berøres av størrelsen på den største observasjonen, får vi ikke signifikant forskjell mellom fordelingene i de to kategoriene. På

den annen side er medianen lik 80 tusen i kategori 3 og 65 tusen i kategori 4. Selv om denne forskjellen ikke er signifikant, kan den indikere en mulig forskjell mellom fordelingene, men evidensen for dette er tynn.

Om vi ønsker en prediksjon av forventning og median for ikke-nyregistrerte virksomheter i gruppe A, er det således bedre å slå sammen kategori 3 og 4. Resultatet er gitt i tabell 5.15. For andre grupper enn gruppe A er det ikke funnet grunnlag i data for fornuftige prediksjoner.

Tabell 5.15 Forventning og median for endringsbeløp gitt endring ("endret nettoinntekt" av andre typer enn "påvist uteholdt omsetning") for ikke-nyregistrerte virksomheter i gruppe A. Bootstrap (BCa) konfidensgrenser basert på 4000 replikasjoner. Antall i parentes.

			95% Konf. grenser	
	Observert (1000 kr.)	Estimert (1000 kr.)	Nedre	Øvre
Forventning	115 (11)	115	53	341
Median	65 (11)	65	45	102

6 Noen konklusjoner

- Det ble ikke funnet evidens for at screeningen hadde noen effekt for endringstilfeller av typen *" endring av nettoinntekt bortsett fra feilperiodiseringer og feil bruk av mva-satser"* eller *"påvist uteholdt omsetning"*, verken når det gjelder sannsynligheten for avdekking av endringstilfeller eller for endringsbeløpets størrelse gitt endring. Når det gjelder endringer av typen *"endring av merverdiavgift relatert til avgiftsfeil på salgsområdet (uten økning i nettoinntekt)"*, ble det funnet evidens for at screeningen hadde effekt på sannsynligheten for avdekking av endring, men ikke på endringsbeløpets størrelse gitt endring.
- Screeningen kan ha hatt mindre effekt for de nye bransjene i 2007-dataene enn for de tre bransjene i 2006-dataene. For 2006-dataene ble det avdekket 26.5% endringstilfeller i trinn-2-utvalget mens det for 2007-dataene ble avdekket 16.7% endringstilfeller (type 1 eller 2) på trinn 2. Dette kan skyldes forskjeller i forekomsten av endringstilfeller mellom de to (disjunkte) bransje-settene for 2006- og 2007-dataene, men det foreligger også en mulighet at utvidelsen av screeningskriteriet, som ble foretatt før innsamlingen av 2007-dataene, har vært for liberal slik at det har skjedd en utvanning av kriteriet og således bidratt til forskjellen i avdekkingsprosent. Siden de to bransje-settene ikke har noen bransjer felles, inneholder dataene ikke informasjon til å kunne teste denne muligheten.
- Et slående funn var at nesten alle de observerte endringstilfellene forekom i gruppen av virksomheter som består av enkeltmannsvirksomheter uten ansatte (kalt *gruppe A* i rapporten og som utgjør 62% av trinn-1-utvalget). Bare 3 av 24 tilfeller av endring av

nettoinntekt og bare 2 av 16 tilfeller av endring av mva ble avdekket utenfor gruppe A. Av de 32 observerte endringstilfellene i alt var det kun 4 utenfor gruppe A. Siden informasjonsgrunnlaget utenfor gruppe A således er nokså tynt, fører dette til at sannsynlighetsberegninger og predikerte endringsbeløp i rapporten, først og fremst er relevant for gruppe A, og mer usikkert utenfor gruppe A.. Funnet tilsier også at man burde tatt med en oppdatert analyse innenfor gruppe A alene, noe som ikke er gjennomført på grunn av tidsbegrensninger. På den annen side er det ikke sannsynlig at de fire virksomhetene med endringer utenfor gruppe A har influert beregningene innenfor gruppe A vesentlig.

- 70-80% av virksomhetene i hver av de tre bransjene "Snekker", "Fotograf" og "Design" tilhører gruppe A, mens bare 9% av virksomhetene i "Jernvarehandel" tilhører gruppe A.
- Konsentrasjonen om gruppe A, små datasett og tilstedeværelsen av noen ekstreme observasjoner blant endringsbeløpene førte til en betydelig grad av overtilpasning for aktuelle prediksjonsmodeller for endringsbeløp gitt endring. Dette ga seg blant annet utslag i at estimatene for forventet endringsbeløp i alle grupper ble lik de observerte gjennomsnittene i gruppene. Dermed er det lite overføring av informasjon mellom grupper slik at prediksjon av forventet endringsbeløp har liten mening i grupper med få eller ingen observasjoner.
- Data for endringsbeløpenes størrelse gitt endring synes å være karakterisert av en tendens til at de fleste observasjonene ligger på et visst moderat nivå, men at det av og til dukker opp en og annen ekstrem observasjon[4]. Denne tendensen var tydelig i de foreliggende data og kunne også observeres i 2006-dataene fra andre bransjer. Det virker altså som om tendensen er karakteristisk for data av denne typen. Hvis dette er riktig, vil en situasjon som her, der de estimerte forventete endringsbeløpene gitt endring gjennomgående blir lik de observerte gjennomsnittene, føre til underestimering i subgrupper som har relativt få observasjoner og uten ekstreme – og overestimering i subgrupper med relativt få observasjoner der det forekommer ekstreme observasjoner. Om man for eksempel ønsker aggregeringsestimater av endringsbeløp over strata, viser analysen at man helst ikke bør gå utenfor gruppe A. Innenfor gruppe A bør man heller ikke aggregere estimater innenfor subgrupper med for få observasjoner. I en situasjon der det er overføring av informasjon mellom subgrupper (som gjerne viser seg ved at estimatene for forventet endring gitt endring framstår som en utglatting av forskjellige observerte gjennomsnitt), vil situasjonen være litt bedre i og med at flere observasjoner dermed står bak hver prediksjon.
- I lys av tendensen til underestimering og overestimering i subgrupper på grunn av noen få ekstreme observasjoner i små datasett, synes den mest rimelige estimeringen av forventet endringsbeløp gitt endring å gi felles estimat for alle grupper slått sammen som om fordelingen er den samme i alle grupper. Dette gir estimatene i tabell 6.1 der bootstrap standardfeil og konfidensintervall (BCa) også er angitt.

[4] Et grovt anslag basert på data innsamlet til nå, er at opptil ca 10% av endringsbeløpene kan karakteriseres som ekstreme. Både denne rapporten og rapporten for 2006-dataene viser at slike ekstreme observasjoner kan ha stor innflytelse på valg av prediksjonsmodeller og på prediksjoner av endringsbeløps størrelse.

Tabell 6.1 **Estimert forventet endringsbeløp gitt endring felles over alle grupper. Bootstrap standardfeil og konfidensgrenser (4000 replikasjoner). Antall observasjoner i parentes.**

			95% konfidensintervall	
Kategori endringsbeløp	Estimert forventet endringsbeløp gitt endring (1000 kr)	Standard-Feil (1000 kr	Nedre konf.grense	Øvre konf.grense
Endret nettoinntekt	114 (24)	32	69	213
Påvist uteholdt omsetning	57 (16)	17	31	99
Andre typer endret nettoinntekt enn påvist uteholdt omsetning	131 (14)	51	55	287
Endring av merverdiavgift relatert til avgiftsfeil på salgsområdet (uten økning i nettoinntekt)	62 (13)	44	14	245

- Tendensen til at det dukker opp sporadiske ekstreme observasjoner blant avdekkete endringsbeløp er et interessant statistisk funn (også diskutert i avsnitt 6.3 i SR) som bør ha konsekvenser i senere studier for modellering av endringsbeløpets størrelse gitt endring. Klassen av gammafordelinger som er brukt som grunnlag for analysen både i FR og her, kan i noen grad fange opp en slik tendens, men, som det er grunn til å tro, i utilstrekkelig grad for data av typen endringsbeløp som hos oss (jfr. for eksempel figur 2 i avsnitt 5.1.2).
- Når det gjelder sannsynligheten for avdekking, kommer ikke overtilpasning inn på samme måte som for endringsbeløpenes størrelse. For det første er datagrunnlaget større (191 observasjoner på trinn 2), og for det andre er eksakt prediksjon ikke aktuelt i en logistisk regresjon siden eksakt prediksjon i data ville implisere noen regresjonskoeffisienter lik pluss eller minus uendelig, og slike tilfeller må lukes ut før en regresjon kan kjøres.
- Med de forbehold som er formulert på slutten av avsnitt 1 er det evidens for at sannsynligheten for avdekking av "endring av nettoinntekt" er høyere i Skatt øst enn andre regioner. Likeledes høyere i mest sentrale kommuner sammenlignet med andre kommuner (se tabell 4.11, 4.12). Det er også betydelig forskjell mellom enkeltmannsvirksomheter og andre virksomheter, der enkeltmannsvirksomheter har størst sannsynlighet. Kombinasjonen med høyest avdekkingssannsynlighet (0.23) er enkeltmannsvirksomheter fra de mest sentrale kommunene, og den minst sannsynlige kombinasjonen er ikke-enkeltmannsvirksomheter (dvs. hovedsakelig AS) fra mindre sentrale kommuner.

- Det er grunn til å tro at responskategorien "endring av nettoinntekt" kan være noe for grov eller inhomogen, siden den omfatter forskjellige årsaker som virker forskjellig, for eksempel "påvist uteholdt omsetning" eller andre årsaker. Et utslag av dette er at sannsynligheten for "endring av nettoinntekt" synes ifølge tabell 4.1 bare å avhenge av virksomhetstype, region og kommunesentralitet, mens sannsynligheten for "påvist uteholdt omsetning" i tillegg avhenger av omsetning og om virksomheten har ekstern regnskapsfører eller ikke (tabell 4.14). Det at de to siste variablene forsvinner fra forklaringen av sannsynligheten for "endring av nettoinntekt", kan være en indikasjon på at kategorien "endring av nettoinntekt" er uhensiktsmessig som responskategori og bør splittes opp.
- Når det gjelder forskjeller i effekter mellom regioner, kom det fram på høringen av en tidligere versjon av denne rapporten, at det har vært forskjellige kriterier ved skattekontorene for utvelging av inspektører spesielt på trinn 1. Hvis således erfaringsgrunnlaget for inspektørene har vært systematisk forskjellige, kan dette ha påvirket sannsynligheten på funn på screeningen på trinn 1 og således på effekten av screeningen. Det er også dermed uklart hvor mye påviste forskjeller mellom regioner indikert av dataanalysen skyldes reelle forskjeller mellom regioner og hvor mye skyldes forskjellig praksis og erfaringsbakgrunn blant inspektørene. Ved framtidige undersøkelser bør det derfor legges vekt på at skattekontorene velger felles kriterier for utvalg av inspektører.
- Når det gjelder avdekking av typen " *endring av merverdiavgift relatert til avgiftsfeil på salgsområdet (uten økning i nettoinntekt)*", er de estimerte sannsynlighetene høyest i Skatt Midt-Norge, noe lavere i Skatt Øst og lavest i Sør, Vest og Nord (jfr. tabell 4.13a,b,c). Virksomheter uten ekstern regnskapsfører ligger litt høyere i sannsynlighet (gjennomsnittlig rundt 20% høyere) enn de med. Kombinasjoner med høyest estimert sannsynlighet i alle regioner er virksomheter uten ekstern regnskapsfører og fra de mest sentrale kommunene.
- Sannsynligheten for avdekking av "*påvist uteholdt omsetning*" synes å være høyest blant virksomheter fra mest sentrale kommuner, uten ekstern regnskapsfører og som har en omsetning under en million (jfr. tabell 4.15). Den laveste sannsynligheten ble estimert for virksomheter utenfor mest sentrale kommuner, med ekstern regnskapsfører og omsetning over en million.
- Sannsynligheten for avdekking av "*endring av nettoinntekt av andre typer enn påvist uteholdt omsetning*" synes å være høyest blant enkeltmannsvirksomheter i Skatt Øst med omsetning over en million (sannsynlighet 0.44 med 0.20 som ensidig nedre konfidensgrense). Den minste sannsynligheten ble estimert for ikke-enkeltmannsvirksomheter utenfor Skatt Øst med omsetning under en million. I tillegg er det verdt å nevne at alle virksomhetene i materialet med denne typen avdekking, bortsett fra en, var ikke-nyregistrert (4 år eller eldre), og de fleste (11/14) hadde ekstern regnskapsfører.
- Den foreliggende analysen har avdekket to potensielt interessante oppgaver som eventuelt kunne være tema for en supplerende rapport senere. På grunn av tidsbegrensninger er de ikke systematisk behandlet i denne rapporten. Den første oppgaven er å gjennomføre en systematisk studie av gruppe A (enkeltmannsvirksomheter uten ansatte) alene. Siden 88% av alle avdekninger forekom i gruppe A, er de fleste resultater og tendenser rapportert her antakelig mest relevant for gruppe A. Jeg gjennomførte noen prøveberegninger i flere av analysene der jeg erstattet (dummy)variabelen for enkeltmannsvirksomhet med en dummyvariabel for gruppe A og fant kun neglisjerbare endringer i estimerte sannsynligheter og forventete endringsbeløp. I tillegg hadde gruppe A signifikant

betydning i alle tilfeller der variabelen for enkeltmannsvirksomhet var signifikant. Den dikotome variabelen *Gruppe A* er et eksempel på en samspillsvariabel[5] som har betydning. Den andre oppgaven ville være å gjennomføre en mer systematisk søking etter andre samspillsvariable som kan være viktige.

Appendiks 1: Simultanfordelingen for indikatorene for "endret nettoinntekt" og "påvist uteholdt omsetning" fra avsnitt 4.7

En mulighet for å unngå selvmotsigelser som i tabell 4.16, er å modellere den simultane fordelingen for Y_1 og Y_3. Den simultane fordelingen for "endret nettoinntekt" (Y_1) og "påvist uteholdt omsetning" (Y_3) er gitt ved

$$f_U(y_1, y_3) = P(Y_1 = y_1 \cap Y_3 = y_3 \mid U) \quad \text{for} \quad y_1, y_3 = 0 \text{ eller } 1.$$

Vi får dermed ved multiplikasjons-setningen

$$\begin{aligned} f_U(0,0) &= P(Y_3 = 0 \mid U) \cdot P(Y_1 = 0 \mid U, Y_3 = 0) \\ f_U(1,0) &= P(Y_3 = 0 \mid U) \cdot P(Y_1 = 1 \mid U, Y_3 = 0) \\ f_U(0,1) &= 0 \\ f_U(1,1) &= P(Y_3 = 1 \mid U) \end{aligned}$$

Skriver vi kort $p_3(U) = P(Y_3 = 1 \mid U)$ og $p_{1|3}(U) = P(Y_1 = 1 \mid U, Y_3 = 0)$, får vi

$$\begin{aligned} f_U(0,0) &= (1 - p_3(U)) \cdot (1 - p_{1|3}(U)) \\ f_U(1,0) &= (1 - p_3(U)) \cdot p_{1|3}(U) \\ f_U(0,1) &= 0 \\ f_U(1,1) &= p_3(U) \end{aligned} \tag{4.7.2}$$

Av denne modellen følger at

$$P(Y_1 = 1 \mid U) = p_3(U) + (1 - p_3(U)) \cdot p_{1|3}(U) \tag{4.7.3}$$

og vi ser at (4.7.1) er automatisk oppfylt.

[5] Kan tolkes som et samspill ved at dummy-variabelen for GruppeA er produktet av dummy-variablene for enkeltmannsvirksomhet og ingen ansatte.

Vi antar videre vanlige logistiske regresjonsmodeller for $p_3(U) = P(Y_3 = 1 | U)$ og $p_{1|3}(U) = P(Y_1 = 1 | U, Y_3 = 0)$. Merk at denne modellen er litt forskjellig fra de logistiske regresjonsmodellene brukt ovenfor. Den impliserer at den marginale modellen for Y_3 er den samme logistiske som ligger bak tabell 4.8 og 4.15, som betyr at $p_3(U)$ allerede er analysert og estimert i avsnitt 4.3 og 4.6. Men den marginale modellen for Y_1 er ikke lenger en vanlig logistisk regresjonsmodell som den som ligger bak tabell 4.1, snarere en blanding av to logistiske regresjonsmodeller.

Det som gjenstår for å estimere f_U, er således å estimere $p_{1|3}(U) = P(Y_1 = 1 | U, Y_3 = 0)$. Vi bruker samme metodikk som før - nemlig ved først å ta utgangspunkt i en estimerbar full modell og deretter på bakgrunn av denne å velge en rimelig prediksjonsmodell som grunnlag for å beregne estimerte sannsynligheter. Selv om screeningsvariabelen *Z* ikke syntes å ha betydning for fordelingene for Y_1 og Y_3 i avsnitt 4.1 og 4.3, så kan det likevel være mulig at den dukker opp i prediksjonsmodellen for $Y_1 | (Y_3 = 0)$. Det er derfor viktig at *Z* er med i den fulle modellen. $Y_3 = 0$ - gruppen består av 175 observasjoner (172 observasjoner med omsetning) på trinn 2, hvorav 8 har $Y_1 = 1$.

Tabell 4.17 Regresjonsresultater (logistisk regresjon) for avdekking av typen "endret nettoinntekt" gitt "ikke uteholdt omsetning".

(Basert på utskrift Ut11 og Ut12 i appendiks A5)

Avhengig Y_1 gitt $Y_3 = 0$	Full modell		Prediksjonsmodell	
	Koeff.	p-verdi	Koeff.	p-verdi
AS	-----	-----	-----	-----
ENK	1.8494	0.266	-----	-----
Ost	4.0376	0.026	2.1566	0.005
Sor	0.6422	0.690	-----	-----
Vest	-----	-----	-----	-----
Midt	0.2364	0.878	-----	-----
Nord	-----	-----	-----	-----
Snekker	1.1070	0.415	-----	-----
Jernv	-0.3372	0.849	-----	-----
Fotograf	-----	-----	-----	-----
Design	-----	-----	-----	-----
Z	-2.1909	0.109	-----	-----
Nyreg	0.3469	0.816	-----	-----
R	-----	-----	-----	-----
Oms0_3	-0.2780	0.834	-----	-----
Oms3_10	-2.2769	0.070	-----	-----
Oms0_10	-----	-----	-----	-----
A0	-0.1885	0.918	-----	-----
A1	0.4326	0.815	-----	-----
Komtjenest	0.5856	0.616	-----	-----
Komsentral	-1.4395	0.318	-----	-----
konstant	-4.7488	0.042	-3.8430	0.000

Antall obs	172		175	172
Log-likelihood	-22.1215		-28.4298	-28.0566
-2 log LR				11.8700
P-verdi redusert vs full modell				0.538

Som før skyldes hullene i den fulle modellen enten eksakt kollinearitet eller utilstrekkelig informasjon i data for å oppnå estimater med rimelig presisjon.

Tabell 4.18 viser p-verdier og informasjonsverdier for noen konkurrerende nærliggende prediksjonsmodeller (PM), med og uten *Z*.

Tabell 4.18 P-verdier og informasjonsverdier for noen prediksjonsmodeller (PM). Forklaringsvariable som inngår i en PM framgår av hvilke p-verdier som er angitt.

PM	*ENK*	*Øst*	*Oms0_10*	*Z*	AIC	BIC
1	0.115	0.008	0.024	---	58.0	70.6
2	0.081	0.005	0.077	0.243	58.6	74.3
3	---	0.002	0.096	---	59.2	68.8
4	---	0.002	0.207	0.423	60.5	73.1
5	---	0.005	---	---	60.9	67.2
6	---	0.002	---	0.197	61.1	70.6

Valget mellom disse 6 PM-ene synes å stå mellom PM1 og PM5, der AIC foretrekker PM1, og BIC PM5. Jeg valgte her PM5. Øst er en klar kandidat og bør være med, mens *ENK* og *Oms0_10* ligger litt på grensen. I lys av det relativt svake datagrunnlaget for Y_1 i dette subsettet av data, valgte jeg ikke å ta dem med.

Appendiks 2: Utskrifter

(Hovedsakelig basert på STATA)

I Utskrifter for Y_1 (indikator for treff av typen "endret nettoinntekt")

Ut1 Y1: Full modell

```
Logistic regression                                     Number of obs   =        188
                                                        LR chi2(18)     =      24.41
                                                        Prob > chi2     =     0.1419
Log likelihood = -59.592586                             Pseudo R2       =     0.1700

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          AS |  -2.422027   2.081755    -1.16   0.245    -6.502192    1.658138
         ENK |   .3577436   1.233761     0.29   0.772    -2.060384    2.775871
         Ost |   1.368574    1.00173     1.37   0.172    -.5947813    3.331929
         Sor |   .7033481    1.04159     0.68   0.500    -1.338131    2.744827
        Vest |   .0395097   1.151994     0.03   0.973    -2.218358    2.297377
        Midt |   1.150024   .9524142     1.21   0.227    -.7166739    3.016721
     Snekker |    .569744   .7747448     0.74   0.462    -.9487279    2.088216
       Jernv |   1.157411   1.152984     1.00   0.315    -1.102397    3.417219
    Fotograf |   .7514232   .9026032     0.83   0.405    -1.017647    2.520493
           Z |   .0984589   .5837303     0.17   0.866    -1.045631    1.242549
       Nyreg |  -.2229266   .7798954    -0.29   0.775    -1.751494     1.30564
           R |   -.463912   .5485793    -0.85   0.398    -1.539108    .6112837
      Oms0_3 |   .3959031   .7540675     0.53   0.600    -1.082042    1.873848
     Oms3_10 |  -.4484624   .7017928    -0.64   0.523    -1.823951    .9270263
          A0 |  -.2055981   1.740918    -0.12   0.906    -3.617735    3.206539
          A1 |  -.2045501   1.712721    -0.12   0.905    -3.561422    3.152322
  Komtjenest |   .1118189   .6302659     0.18   0.859     -1.12348    1.347117
  Komsentral |   .7259465   .6747805     1.08   0.282     -.596599    2.048492
       _cons |  -3.220979   2.390432    -1.35   0.178     -7.90614    1.464183
------------------------------------------------------------------------------
```

Ut2 Y1: Redusert modell 1

```
Logistic regression                                     Number of obs   =        191
                                                        LR chi2(2)      =      14.50
                                                        Prob > chi2     =     0.0007
Log likelihood = -64.958124                             Pseudo R2       =     0.1004

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Ost |   1.048143   .4730356     2.22   0.027     .1210103    1.975276
          AS |  -2.086303   1.043592    -2.00   0.046    -4.131706   -.0408995
       _cons |  -1.944197   .2934143    -6.63   0.000    -2.519278   -1.369116
------------------------------------------------------------------------------
```

Ut3 Y1: Redusert modell 2

```
Logistic regression                                     Number of obs   =        191
                                                        LR chi2(2)      =      14.19
                                                        Prob > chi2     =     0.0008
Log likelihood = -65.110973                             Pseudo R2       =     0.0983

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Komsentral |   .9744545   .4838599     2.01   0.044     .0261067    1.922802
          AS |  -2.129327   1.041003    -2.05   0.041    -4.169656    -.088998
       _cons |  -2.208121   .4039965    -5.47   0.000     -2.99994   -1.416303
------------------------------------------------------------------------------
```

II Utskrifter for Y_2 (indikator for treff av typen " endring merverdiavgift - kun relatert til avgiftsfeil på salgsområdet (uten økning i nettoinntekt)")

Ut4 Y2: Full modell

```
Logistic regression                                     Number of obs   =        188
                                                        LR chi2(14)     =      22.08
                                                        Prob > chi2     =     0.0770
Log likelihood = -38.789237                             Pseudo R2       =     0.2215

------------------------------------------------------------------------------
          Y2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          AS |  -1.542195   1.230673    -1.25   0.210    -3.954271    .8698805
         Ost |   .3011016   .8510947     0.35   0.724    -1.367013    1.969217
        Midt |    1.93118   .9012663     2.14   0.032     .1647302    3.697629
     Snekker |   .4652664     1.3145     0.35   0.723    -2.111107     3.04164
    Fotograf |  -1.074357   1.805909    -0.59   0.552    -4.613874     2.46516
      Design |   .7243252   1.497226     0.48   0.629    -2.210185    3.658835
           Z |   1.943452   .8817412     2.20   0.028      .215271    3.671633
       Nyreg |   .3359929    .959525     0.35   0.726    -1.544642    2.216627
           R |   -.541425   .7106156    -0.76   0.446    -1.934206    .8513559
      Oms0_3 |  -.8765083   .9200836    -0.95   0.341    -2.679839    .9268224
     Oms3_10 |  -1.433739   .9069688    -1.58   0.114    -3.211365    .3438873
          A1 |   1.758289   1.162512     1.51   0.130     -.520193    4.036771
  Komtjenest |   .4600108   .7604267     0.60   0.545    -1.030398     1.95042
  Komsentral |   1.187585   .9070052     1.31   0.190    -.5901126    2.965282
       _cons |  -4.424464   1.508548    -2.93   0.003    -7.381163   -1.467765
------------------------------------------------------------------------------
```

Ut5 Y2: Redusert modell

```
Logistic regression                                     Number of obs   =        191
                                                        LR chi2(3)      =      13.64
                                                        Prob > chi2     =     0.0034
Log likelihood = -43.238025                             Pseudo R2       =     0.1363

------------------------------------------------------------------------------
          Y2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Komsentral |   1.218789   .6400607     1.90   0.057    -.0357074    2.473284
        Midt |   1.181717      .6972     1.69   0.090    -.1847701    2.548204
           Z |   1.847009   .7019175     2.63   0.009     .4712759    3.222742
       _cons |  -4.656804   .8242416    -5.65   0.000    -6.272288    -3.04132
------------------------------------------------------------------------------
```

III Utskrifter for Y_3 (indikator for treff av typen " Konkret påvist uteholdt omsetning (kontrollmelding o.l på SALG som ikke er bokført)")

Ut6 Y3: Full modell

```
Logistic regression                                     Number of obs   =        188
                                                        LR chi2(15)     =      26.05
                                                        Prob > chi2     =     0.0375
Log likelihood = -41.693558                             Pseudo R2       =     0.2381

------------------------------------------------------------------------------
          Y3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |   1.890222   1.392653     1.36   0.175    -.8393268    4.619772
```

```
         Ost |   .2831986   1.103601     0.26   0.797     -1.87982    2.446217
         Sor |   .6099199    1.13434     0.54   0.591    -1.613346    2.833186
        Vest |  -.6482109   1.422752    -0.46   0.649    -3.436753    2.140331
        Midt |   .9519314   1.042249     0.91   0.361     -1.09084    2.994702
     Snekker |    .047878   .9622227     0.05   0.960    -1.838044      1.9338
       Jernv |   1.332818    1.49029     0.89   0.371    -1.588097    4.253732
    Fotograf |   .7780766   1.041681     0.75   0.455    -1.263581    2.819734
           Z |   .5194499   .7346561     0.71   0.480    -.9204496    1.959349
       Nyreg |  -.2196021   .9260023    -0.24   0.813    -2.034533    1.595329
           R |  -1.039187   .6218608    -1.67   0.095    -2.258012    .1796375
      Oms0_3 |    1.70579   1.241243     1.37   0.169    -.7270023    4.138582
     Oms3_10 |   1.407201   1.239224     1.14   0.256    -1.021634    3.836037
   Komtjenest |   .0991412   .7636371     0.13   0.897     -1.39756    1.595842
   Komsentral |   1.187279   .7950838     1.49   0.135    -.3710563    2.745615
        _cons |  -6.313211   2.252024    -2.80   0.005     -10.7271   -1.899324
------------------------------------------------------------------------------
```

Ut7 Y3: Redusert modell 1

```
Logistic regression                               Number of obs   =        191
                                                  LR chi2(3)      =      15.04
                                                  Prob > chi2     =     0.0018
Log likelihood = -47.464065                       Pseudo R2       =     0.1368

------------------------------------------------------------------------------
          Y3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Komsentral |   1.329956   .6123994     2.17   0.030     .1296755    2.530237
           R |  -1.260047   .5510521    -2.29   0.022    -2.340089   -.1800048
           Z |   1.003809   .5783709     1.74   0.083    -.1297771    2.137395
       _cons |  -2.998568   .6905991    -4.34   0.000    -4.352118   -1.645019
------------------------------------------------------------------------------
```

Ut8 Y3: Redusert modell 2

```
Logistic regression                               Number of obs   =        188
                                                  LR chi2(3)      =      18.24
                                                  Prob > chi2     =     0.0004
Log likelihood = -45.602459                       Pseudo R2       =     0.1666

------------------------------------------------------------------------------
          Y3 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Komsentral |   1.229153   .6159337     2.00   0.046     .0219452    2.436361
           R |  -1.154529   .5558675    -2.08   0.038    -2.244009   -.0650489
      Oms0_10 |   2.027235   1.056566     1.92   0.055    -.0435968    4.098067
       _cons |  -4.055696   1.133351    -3.58   0.000    -6.277024   -1.834368
------------------------------------------------------------------------------
```

IV Utskrifter for funn på trinn 1 (*Z*)

Ut 9 Z: Full modell

```
Logistic regression                               Number of obs   =        456
                                                  LR chi2(16)     =      56.35
                                                  Prob > chi2     =     0.0000
Log likelihood =  -198.3359                       Pseudo R2       =     0.1244

------------------------------------------------------------------------------
          Z |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ENK |   .7007879   .5203982     1.35   0.178    -.3191738     1.72075
        Ost |   1.455535   .4628673     3.14   0.002     .5483318    2.362738
        Sor |   .3839864    .442792     0.87   0.386      -.48387    1.251843
       Vest |   .3194487   .4675385     0.68   0.494    -.5969099    1.235807
       Midt |  -.2822853   .4540061    -0.62   0.534    -1.172121    .6075504
    Snekker |   .2172646   .3963981     0.55   0.584    -.5596614    .9941906
      Jernv |   .0722427   .6055596     0.12   0.905    -1.114632    1.259118
```

```
    Fotograf |   .2905305   .4627525     0.63   0.530    -.6164477    1.197509
       Nyreg |  -.2169752   .4199148    -0.52   0.605    -1.039993    .6060428
          A0 |   .8444727   .7207515     1.17   0.241    -.5681743     2.25712
          A1 |   .4182271   .6615519     0.63   0.527    -.8783908    1.714845
  Komtjenest |   .6604845   .3027501     2.18   0.029     .0671052    1.253864
  Komsentral |   -.627048   .3476962    -1.80   0.071     -1.30852    .0544241
           R |   -.691459   .2895468    -2.39   0.017     -1.25896   -.1239576
       Oms03 |   .3019953   .4387867     0.69   0.491    -.5580109    1.162001
      Oms310 |   .3746075   .4013621     0.93   0.351    -.4120478    1.161263
       _cons |  -2.909784   .8076242    -3.60   0.000    -4.492699    -1.32687
------------------------------------------------------------------------------
```

Ut 10 Z: Redusert modell

```
Logistic regression                               Number of obs   =        467
                                                  LR chi2(4)      =      44.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -209.57049                       Pseudo R2       =     0.0956

------------------------------------------------------------------------------
           Z |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |    1.391903   .3275816     4.25   0.000     .7498552    2.033951
         Ost |    1.144231   .2841636     4.03   0.000     .5872804    1.701181
    KSminKTJ |   -.5674579   .2196798    -2.58   0.010    -.9980224   -.1368934
           R |    -.492227   .2680243    -1.84   0.066    -1.017545    .0330909
       _cons |   -2.347143   .3530828    -6.65   0.000    -3.039172   -1.655113
------------------------------------------------------------------------------
```

V Utskrifter for kombinert analyse for Y1 og Y3

Ut 11 Y1 | Y3 = 0: Full modell

```
Logistic regression                               Number of obs   =        172
                                                  LR chi2(14)     =      20.47
                                                  Prob > chi2     =     0.1161
Log likelihood = -22.121549                       Pseudo R2       =     0.3163

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |    1.849433   1.661183     1.11   0.266    -1.406427    5.105293
         Ost |    4.037587   1.814561     2.23   0.026     .4811121    7.594062
         Sor |    .6422427   1.609291     0.40   0.690     -2.51191    3.796395
        Midt |    .2364377   1.539848     0.15   0.878    -2.781609    3.254485
     Snekker |    1.107003   1.359298     0.81   0.415    -1.557172    3.771177
       Jernv |   -.3372259   1.776882    -0.19   0.849    -3.819852      3.1454
           Z |   -2.190879   1.365695    -1.60   0.109    -4.867593    .4858347
       Nyreg |    .3469144   1.491553     0.23   0.816    -2.576477    3.270305
      Oms0_3 |   -.2780113   1.328533    -0.21   0.834    -2.881889    2.325866
     Oms3_10 |   -2.276948   1.255383    -1.81   0.070    -4.737453    .1835565
          A0 |   -.1885225   1.822271    -0.10   0.918    -3.760108    3.383063
          A1 |    .4325543   1.849431     0.23   0.815    -3.192263    4.057372
  Komtjenest |    .5855733   1.167164     0.50   0.616    -1.702027    2.873173
  Komsentral |   -1.439536   1.442564    -1.00   0.318    -4.266909    1.387837
       _cons |    -4.74876   2.331276    -2.04   0.042    -9.317976   -.1795438
------------------------------------------------------------------------------
```

Ut 12 Y1 | Y3 = 0: Redusert modell

```
Logistic regression                                     Number of obs   =        175
                                                        LR chi2(1)      =       8.13
                                                        Prob > chi2     =     0.0043
Log likelihood =  -28.42977                             Pseudo R2       =     0.1252

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Ost |    2.156631   .7599309     2.84   0.005     .6671939    3.646068
       _cons |    -3.84303   .5835034    -6.59   0.000    -4.986676   -2.699385
------------------------------------------------------------------------------
```

Ut 13 Y1a: Full modell

```
Logistic regression                                     Number of obs   =        172
                                                        LR chi2(14)     =      20.47
                                                        Prob > chi2     =     0.1161
Log likelihood = -22.121549                             Pseudo R2       =     0.3163

------------------------------------------------------------------------------
         Y1a |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |    1.849433   1.661183     1.11   0.266    -1.406427    5.105293
         Ost |    4.037587   1.814561     2.23   0.026     .4811121    7.594062
         Sor |    .6422427   1.609291     0.40   0.690     -2.51191    3.796395
        Midt |    .2364377   1.539848     0.15   0.878    -2.781609    3.254485
     Snekker |    1.107003   1.359298     0.81   0.415    -1.557172    3.771177
       Jernv |   -.3372259   1.776882    -0.19   0.849    -3.819852      3.1454
           Z |   -2.190879   1.365695    -1.60   0.109    -4.867593    .4858347
       Nyreg |    .3469144   1.491553     0.23   0.816    -2.576477    3.270305
     Oms0_3  |   -.2780113   1.328533    -0.21   0.834    -2.881889    2.325866
    Oms3_10  |   -2.276948   1.255383    -1.81   0.070    -4.737453    .1835565
          A0 |   -.1885225   1.822271    -0.10   0.918    -3.760108    3.383063
          A1 |    .4325543   1.849431     0.23   0.815    -3.192263    4.057372
  Komtjenest |    .5855733   1.167164     0.50   0.616    -1.702027    2.873173
  Komsentral |   -1.439536   1.442564    -1.00   0.318    -4.266909    1.387837
       _cons |    -4.74876   2.331276    -2.04   0.042    -9.317976   -.1795438
------------------------------------------------------------------------------
```

Ut 14 Y1a: Redusert modell

```
Logistic regression                                     Number of obs   =        172
                                                        LR chi2(3)      =      14.69
                                                        Prob > chi2     =     0.0021
Log likelihood = -25.009481                             Pseudo R2       =     0.2270

------------------------------------------------------------------------------
         Y1a |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |    1.857461   1.177797     1.58   0.115    -.4509776     4.1659
         Ost |    2.158824   .8119007     2.66   0.008     .5675282     3.75012
     Oms0_10 |   -1.903153    .843707    -2.26   0.024    -3.556789    -.249518
       _cons |   -4.256336   1.060103    -4.02   0.000    -6.334099   -2.178573
------------------------------------------------------------------------------
```

VI Utskrifter for separat analyse av X1 (beløp for "endret nettoinntekt")

Ut 15 Full modell for X1

```
Generalized linear models                           No. of obs      =        24
```

```
Optimization     : ML                                Residual df     =         7
                                                     Scale parameter =  .7479106
Deviance         =  11.84237181                      (1/df) Deviance =  1.691767
Pearson          =  5.235374041                      (1/df) Pearson  =  .7479106

Variance function: V(u) = u^2                        [Gamma]
Link function    : g(u) = ln(u)                      [Log]

                                                     AIC             =  25.85141
Log likelihood   = -293.2169571                      BIC             = -10.40401

------------------------------------------------------------------------------
             |                 OIM
          X1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |    2.895075   2.194102     1.32   0.187    -1.405287    7.195436
         Ost |   -.8627141   1.025173    -0.84   0.400    -2.872016    1.146588
         Sor |   -.1342071   1.498970    -0.09   0.929    -3.072134     2.80372
        Vest |    -.941011   1.369318    -0.69   0.492    -3.624825    1.742803
        Midt |    .2167944   1.374656     0.16   0.875    -2.477482    2.911071
     Snekker |   -1.616937   1.600294    -1.01   0.312    -4.753455    1.519581
    Fotograf |   -1.263878   1.312708    -0.96   0.336    -3.836739    1.308982
      Design |   -1.863186   1.748186    -1.07   0.287    -5.289568    1.563195
           Z |    1.692951   .7986682     2.12   0.034     .1275903    3.258312
           R |    3.006117   .7829627     3.84   0.000     1.471538    4.540696
      NyregR |   -2.356687   .9955399    -2.37   0.018     -4.30791   -.4054648
          A0 |   -2.639812   1.824024    -1.45   0.148    -6.214834    .9352098
          A1 |   -.8473384   1.847249    -0.46   0.646    -4.467879    2.773202
  Komtjenest |   -.6001073   1.204895    -0.50   0.618    -2.961657    1.761443
  Komsentral |    .6745015   1.384252     0.49   0.626    -2.038582    3.387585
     Oms0_10 |    1.447407   .7603133     1.90   0.057    -.0427792    2.937594
       _cons |    9.026322   2.139638     4.22   0.000     4.832709    13.21993
------------------------------------------------------------------------------
```

Ut 16 Redusert model 1 for X1

```
Generalized linear models                            No. of obs      =        24
Optimization     : ML                                Residual df     =        16
                                                     Scale parameter =  .5752265
Deviance         =  15.00362043                      (1/df) Deviance =  .9377263
Pearson          =  9.203623368                      (1/df) Pearson  =  .5752265

Variance function: V(u) = u^2                        [Gamma]
Link function    : g(u) = ln(u)                      [Log]

                                                     AIC             =  25.23313
Log likelihood   = -294.7975814                      BIC             = -35.84524

------------------------------------------------------------------------------
             |                 OIM
          X1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |    2.354651   .6298206     3.74   0.000     1.120225    3.589077
     OstVest |   -1.114265   .4515118    -2.47   0.014    -1.999212    -.229318
           Z |    1.158095   .4079569     2.84   0.005     .3585145    1.957676
           R |    2.078037   .5536977     3.75   0.000     .9928095    3.163265
      NyregR |   -2.063346    .655504    -3.15   0.002     -3.34811   -.7785821
          A0 |   -2.958126   .8183844    -3.61   0.000    -4.562129   -1.354122
     Oms0_10 |    1.388752   .5192619     2.67   0.007     .3710175    2.406487
       _cons |    9.635575   .7432116    12.96   0.000     8.178907    11.09224
------------------------------------------------------------------------------
```

Ut 17 Redusert model 2 for X1

```
Generalized linear models                            No. of obs      =        24
Optimization     : ML                                Residual df     =        21
                                                     Scale parameter =  1.360111
Deviance         =  25.57184805                      (1/df) Deviance =  1.217707
Pearson          =  28.56233581                      (1/df) Pearson  =  1.360111

Variance function: V(u) = u^2                        [Gamma]
```

```
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  25.25681
Log likelihood   = -300.0816952                    BIC             = -41.16728

------------------------------------------------------------------------------
             |                 OIM
          X1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |   1.637756    .862432     1.90   0.058    -.0525794    3.328092
           R |   .8284197   .4923618     1.68   0.092    -.1365916    1.793431
       _cons |   9.484365   .8670378    10.94   0.000     7.785002    11.18373
------------------------------------------------------------------------------
```

Ut 18 Redusert model 3 for X1

```
Generalized linear models                          No. of obs      =        24
Optimization     : ML                              Residual df     =        21
                                                   Scale parameter =  1.849832
Deviance         =   24.8739656                    (1/df) Deviance =  1.184475
Pearson          =  38.84647449                    (1/df) Pearson  =  1.849832

Variance function: V(u) = u^2                      [Gamma]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  25.22773
Log likelihood   =  -299.732754                    BIC             = -41.86516

------------------------------------------------------------------------------
             |                 OIM
          X1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |   2.347841    1.07349     2.19   0.029     .2438394    4.451843
          A0 |   -1.38186    1.07349    -1.29   0.198    -3.485862    .7221414
       _cons |   10.60338   1.153516     9.19   0.000     8.342532    12.86423
------------------------------------------------------------------------------
```

VII Utskrifter for analyse av X2 (beløp for “endring av merverdiavgift”)

Ut 19 Full modell for X2

```
Generalized linear models                          No. of obs      =        13
Optimization     : ML                              Residual df     =         1
                                                   Scale parameter =  1.652356
Deviance         =  3.499448181                    (1/df) Deviance =  3.499448
Pearson          =  1.652356212                    (1/df) Pearson  =  1.652356

Variance function: V(u) = u^2                      [Gamma]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =  23.08934
Log pseudolikelihood = -138.0807142                BIC             =  .9344988

------------------------------------------------------------------------------
             |               Robust
          X2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |  -2.298517   .3344821    -6.87   0.000     -2.95409   -1.642944
         Ost |  -.0107469   .3344821    -0.03   0.974    -.6663198     .644826
        Midt |   .9789049   4.42e-08  2.2e+07   0.000     .9789049     .978905
     Snekker |  -.4742363   .3344821    -1.42   0.156    -1.129809    .1813366
       Jernv |  -2.972477   1.003446    -2.96   0.003    -4.939196   -1.005759
           Z |   .1501298   .3344821     0.45   0.654    -.5054431    .8057027
  Komtjenest |  -.4270404   .3344821    -1.28   0.202    -1.082613    .2285325
  Komsentral |   1.512696   .3344821     4.52   0.000     .8571231    2.168269
       Nyreg |   2.425079   2.35e-08  1.0e+08   0.000     2.425079    2.425079
           R |   .0766621   .3344821     0.23   0.819    -.5789108     .732235
     Oms0_10 |  -2.432344   4.79e-07 -5.1e+06   0.000    -2.432345   -2.432343
```

```
       _cons |   12.06961   1.003446    12.03   0.000     10.10289    14.03633
------------------------------------------------------------------------------
```

Ut 20 Redusert modell 1 for X2

```
Generalized linear models                         No. of obs      =         13
Optimization     : ML                             Residual df     =          6
                                                  Scale parameter =  .3184931
Deviance         =  3.992687075                   (1/df) Deviance =  .6654478
Pearson          =  1.910958322                   (1/df) Pearson  =  .3184931

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  22.35805
Log pseudolikelihood = -138.3273336               BIC             = -11.39701

------------------------------------------------------------------------------
             |               Robust
          X2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |  -1.311541   .3830557    -3.42   0.001    -2.062316   -.5607656
       Jernv |  -1.981234   .1291171   -15.34   0.000    -2.234299   -1.728169
  Komtjenest |  -1.094177   .1291171    -8.47   0.000    -1.347242   -.8411118
  Komsentral |   1.821789   .3109951     5.86   0.000     1.21225    2.431328
       Nyreg |   1.911186   .2839624     6.73   0.000     1.35463    2.467742
     Oms0_10 |  -2.270706   .1291171   -17.59   0.000    -2.52377   -2.017641
       _cons |   11.50233   .3109951    36.99   0.000     10.89279    12.11187
------------------------------------------------------------------------------
```

Ut 21 Redusert modell 2 for X2

```
Generalized linear models                         No. of obs      =         13
Optimization     : ML                             Residual df     =          9
                                                  Scale parameter =   .525771
Deviance         =  8.609815274                   (1/df) Deviance =  .9566461
Pearson          =   4.73193931                   (1/df) Pearson  =   .525771

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = ln(u)                   [Log]

                                                  AIC             =  22.25168
Log pseudolikelihood = -140.6358977               BIC             = -14.47473

------------------------------------------------------------------------------
             |               Robust
          X2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |  -2.735998   .7218854    -3.79   0.000    -4.150867   -1.321128
       Nyreg |   2.001282   .2741825     7.30   0.000     1.463894     2.53867
     Oms0_10 |   -1.38251   .3813679    -3.63   0.000    -2.129977   -.6350424
       _cons |   12.67611   .6710181    18.89   0.000     11.36094    13.99128
------------------------------------------------------------------------------
```

VIII Utskrifter for analyse av X3 (beløp for "påvist uteholdt omsetning")

Ut 22 Full modell for X3

```
Generalized linear models                         No. of obs      =         16
Optimization     : ML                             Residual df     =          2
                                                  Scale parameter =  3.314278
Deviance         =  10.22058147                   (1/df) Deviance =  5.110291
Pearson          =  6.628555858                   (1/df) Pearson  =  3.314278
```

```
Variance function: V(u) = u^2                          [Gamma]
Link function    : g(u) = ln(u)                        [Log]

                                                       AIC             =  24.64108
Log pseudolikelihood = -183.1286341                    BIC             =  4.675404
```

X3	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Ost	-.7216087	.2580085	-2.80	0.005	-1.227296	-.2159214
Sor	2.065625	8.177492	0.25	0.801	-13.96196	18.09321
Midt	-1.091365	1.898825	-0.57	0.565	-4.812993	2.630263
Skredder	4.000551	6.61324	0.60	0.545	-8.961161	16.96226
Jernv	-2.685463	8.244457	-0.33	0.745	-18.8443	13.47338
Fotograf	1.804875	2.899045	0.62	0.534	-3.877149	7.486899
Z	-1.467708	6.815115	-0.22	0.829	-14.82509	11.88967
Komtjenest	2.593021	6.999268	0.37	0.711	-11.12529	16.31133
Komsentral	-3.026885	6.729612	-0.45	0.653	-16.21668	10.16291
Nyreg	1.806947	6.841603	0.26	0.792	-11.60235	15.21624
R	-3.397519	13.47933	-0.25	0.801	-29.81652	23.02148
Oms03	.3463679	7.1614	0.05	0.961	-13.68972	14.38245
Oms310	-.5275238	6.832144	-0.08	0.938	-13.91828	12.86323
_cons	11.70049	13.77352	0.85	0.396	-15.29511	38.69609

Ut 23 Redusert modell for X3

```
Generalized linear models                              No. of obs      =        16
Optimization     : ML                                  Residual df     =        13
                                                       Scale parameter =  .6517649
Deviance         =  16.26827847                        (1/df) Deviance =  1.251406
Pearson          =  8.472943063                        (1/df) Pearson  =  .6517649

Variance function: V(u) = u^2                          [Gamma]
Link function    : g(u) = ln(u)                        [Log]

                                                       AIC             =  23.64406
Log pseudolikelihood = -186.1524826                    BIC             = -19.77537
```

X3	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
Ost	-1.420348	.3505191	-4.05	0.000	-2.107353	-.7333436
Jernv	-2.475332	.5428555	-4.56	0.000	-3.53931	-1.411355
_cons	11.38781	.3034243	37.53	0.000	10.7931	11.98251

Frisch Centre Publications

All publications are available in Pdf-format at : www.frisch.uio.no

Rapporter (Reports)

1/2006	**Finansiering av tros- og livssynssamfunn**	Aanund Hylland
2/2006	**Optimale strategier i et to-kvotesystem**	Rolf Golombek, Cathrine Hagem, Michael Hoel
3/2006	**Evaluering av tilskuddsordningen for organisasjoner for personer med nedsatt funksjonsevne**	Rolf Golombek, Jo Thori Lind
4/2006	**Aetats kvalifiserings- og opplæringstiltak – En empirisk analyse av seleksjon og virkninger**	Ines Hardoy, Knut Røed, Tao Zhang
5/2006	**Analyse av aldersdifferensiert arbeidsgiveravgift**	Gaute Ellingsen, Knut Røed
6/2006	**Utfall av yrkesrettet attføring i Norge 1994-2000**	Tyra Ekhaugen
7/2006	**Inntektsfordeling og inntektsmobilitet – pensjonsgivende inntekt i Norge 1971-2003**	Ola Lotherington Vestad
8/2006	**Effektiv måloppnåelse En analyse av utvalgte politiske målsetninger**	Nils-Henrik M. von der Fehr
9/2006	**Sektoranalyser – Gjennomgang av samfunnsøkonomiske analyser av effektiviseringspotensialer for utvalgte sektorer**	Finn R. Førsund
10/2006	**Veien til uføretrygd i Norge**	Elisabeth Fevang, Knut Røed
1/2007	**Generisk bytte En økonometrisk studie av aktørenes og prisenes betydning for substitusjon**	Vivian Almendingen
2/2007	**Firm entry and post-entry performance in selected Norwegian industries**	Ola Lotherington Vestad
1/2008	**Er kommunesektoren og/eller staten lønnsledende? En sammenlikning av lønnsnivå for arbeidstakere i kommunal, statlig og privat sektor**	Elisabeth Fevang, Steinar Strøm, Erik Magnus Sæther
2/2008	**Tjenestepensjon og mobilitet på arbeidsmarkedet**	Nina Skrove Falch
3/2008	**Ressurser i grunnskole og videregående opplæring i Norge 2003-2007**	Torbjørn Hægeland, Lars J. Kirkebøen, Oddbjørn Raaum
4/2008	**Norms and Tax Evasion**	Erling Barth, Alexander W. Cappelen

1/2009	**Revelation of Tax Evasion by Random Audits Report on Main Project, Part 1**	Erling Eide, Harald Goldstein, Paul Gunnar Larssen, Jack-Willy Olsen
2/2009	**Øre for læring – Ressurser i grunnskole og videregående opplæring i Norge 2003-2008**	Torbjørn Hægeland, Lars J. Kirkebøen, Oddbjørn Raaum
3/2009	**Effekter på arbeidstilbudet av pensjonsreformen**	Erik Hernæs, Fedor Iskhakov
1/2010	**Revelation of Tax Evasion by Random Audits. Report on Main Project, Part 2**	Anders Berset, Erling Eide, Harald Goldstein, Paul Gunnar Larssen, Jack-Willy Olsen

Arbeidsnotater (Working papers)

1/2006	**Costs and coverage of occupational pensions**	Erik Hernæs, Tao Zhang
2/2006	**Inntektsfordelingen i Norge, og forskjellige årsaker til ulikheter i pensjonsgivende inntekt**	Ola Lotherington Vestad
3/2006	**The Wage Effect of Computer-use in Norway**	Fitwi H. Wolday
1/2007	**An evaluation of the labour market response of eliminating the retirement earnings test rule**	Erik Hernæs, Zhiyang Jia
1/2008	**LIBEMOD 2000 - LIBeralisation MODel for the European Energy Markets: A Technical Description**	F.R. Aune, K.A. Brekke, R. Golombek, S.A.C. Kittelsen, K.E. Rosendahl
2/2008	**Modelling Households in LIBEMOD 2000 - A Nested CES Utility Function with Endowments**	Sverre Kittelsen
3/2008	**Analyseopplegg for å kunne male om reorganisering av skatteetaten fører til en mer effektiv ressursbruk**	Finn R. Førsund, Sverre A.C. Kittelsen
4/2008	**Patenter i modeler med teknologisk vekst – en litteraturoversikt med vekt på klimapolitikk**	Helge Berglann
5/2008	**The R&D of Norwegian Firms: an Empirical Analysis**	Anton Giulio Manganelli
1/2009	**An Informal Care Leave Arrangement – An Economic Evaluation**	Kebebew Negera

Memoranda (Discussion papers)

The series is published by Department of Economics, University of Oslo, in co-operation with the Frisch Centre. This list includes memoranda related to Frisch Centre projects.
The complete list of memoranda can be found at http://www.oekonomi.uio.no/memo/.

1/2006	**The Determinants of Occupational Pensions**	Erik Hernæs, John Piggott, Tao Zhang and Steinar Strøm
4/2006	**Moving between Welfare Payments. The Case of Sickness Insurance for the Unemployed**	Morten Henningsen
6/2006	**Justifying Functional Forms in Models for Transitions between Discrete States, with Particular Reference to Employment-Unemployment Dynamics**	John Dagsvik
15/2006	**Retirement in Non-Cooperative and Cooperative Families**	Erik Hernæs, Zhiyang Jia, Steinar Strøm
16/2006	**Early Retirement and Company Characteristics**	Erik Hernæs, Fedor Iskhakov and Steinar Strøm
20/2006	**Simulating labor supply behavior when workers have preferences for job opportunities and face nonlinear budget constraints**	John K. Dagsvik, Marilena Locatelli, Steinar Strøm
21/2006	**Climate agreements: emission quotas versus technology policies**	Rolf Golombek, Michael Hoel
22/2006	**The Golden Age of Retirement**	Line Smart Bakken
23/2006	**Advertising as a Distortion of Social Learning**	Kjell Arne Brekke, Mari Rege
24/2006	**Advertising as Distortion of Learning in Markets with Network Externalities**	Kjell Arne Brekke, Mari Rege
26/2006	**Optimal Timing of Environmental Policy; Interaction Between Environmental Taxes and Innovation Externalities**	Reyer Gerlagh, Snorre Kverndokk, Knut Einar Rosendahl
3/2007	**Corporate investment, cash flow level and market imperfections: The case of Norway**	B. Gabriela Mundaca, Kjell Bjørn Nordal
4/2007	**Monitoring, liquidity provision and financial crisis risk**	B. Gabriela Mundaca
5/2007	**Total tax on Labour Income**	Morten Nordberg
6/2007	**Employment behaviour of marginal workers**	Morten Nordberg
9/2007	**As bad as it gets: Well being deprivation of sexually exploited trafficked women**	Di Tommaso M.L., Shima I., Strøm S., Bettio F.
10/2007	**Long-term Outcomes of Vocational Rehabilitation Programs: Labor Market Transitions and Job Durations for Immigrants**	Tyra Ekhaugen
12/2007	**Pension Entitlements and Wealth Accumulation**	Erik Hernæs, Weizhen Zhu

13/2007	**Unemployment Insurance in Welfare States: Soft Constraints and Mild Sanctions**	Knut Røed, Lars Westlie
15/2007	**Farrell Revisited: Visualising the DEA Production Frontier**	Finn R. Førsund, Sverre A. C. Kittelsen, Vladimir E. Krivonozhko
16/2007	**Reluctant Recyclers: Social Interaction in Responsibility Ascription**	Kjell Arne Brekke , Gorm Kipperberg, Karine Nyborg
17/2007	**Marital Sorting, Household Labor Supply, and Intergenerational Earnings Mobility across Countries**	O. Raaum, B. Bratsberg, K. Røed, E. Österbacka, T. Eriksson, M. Jäntti, R. Naylor
18/2007	**Pennies from heaven - Using exogenous tax variation to identify effects of school resources on pupil achievement**	Torbjørn Hægeland, Oddbjørn Raaum and Kjell Gunnar Salvanes
19/2007	**Trade-offs between health and absenteeism in welfare states: striking the balance**	Simen Markussen
1/2008	**Is electricity more important than natural gas? Partial liberalization of the Western European energy markets**	Kjell Arne Brekke, Rolf Golombek, Sverre A.C. Kittelsen
3/2008	**Dynamic programming model of health and retirement**	Fedor Ishakov
8/2008	**Nurses wanted. Is the job too harsh or is the wage too low?**	M. L. Di Tommaso, Steinar Strøm, Erik Magnus Sæther
10/2008	**Linking Environmental and Innovation Policy**	Reyer Gerlagh, Snorre Kverndokk, Knut Einar Rosendahl
11/2008	**Generic substitution**	Kari Furu, Dag Morten Dalen, Marilena Locatelli, Steinar Strøm
14/2008	**Pension Reform in Norway: evidence from a structural dynamic model**	Fedor Iskhakov
15/2008	**I Don't Want to Hear About it: Rational Ignorance among Duty-Oriented Consumers**	Karine Nyborg
21/2008	**Equity and Justice in Global Warming Policy**	Snorre Kverndokk, Adam Rose
22/2008	**The Impact of Labor Market Policies on Job Search Behavior and Post-Unemployment Job Quality**	Simen Gaure, Knut Røed, Lars Westlie
24/2008	**Norwegian Vocational Rehabilitation Programs: Improving Employability and Preventing Disability?**	Lars Westlie
25/2008	**The Long-term Impacts of Vocational Rehabilitation**	Lars Westlie
28/2008	**Climate Change, Catastrophic Risk and the Relative Unimportance of Discounting**	Eric Nævdal, Jon Vislie

29/2008	**Bush meets Hotelling: Effects of improved renewable energy technology on greenhouse gas emissions**	Michael Hoel
7/2009	**The Gate is Open: Primary Care Physicians as Social Security Gatekeepers**	Benedicte Carlsen, Karine Nyborg
9/2009	**Towards an Actuarially Fair Pension System in Norway**	Ugo Colombino, Erik Hernæs, Marilena Locatelli, Steinar Strøm
13/2009	**Moral Concerns on Tradable Pollution Permits in International Environmental Agreements**	Johan Eyckmans, Snorre Kverndokk
14/2009	**Productivity of Tax Offices in Norway**	Finn R. Førsund, Dag Fjeld Edvardsen, Sverre A.C. Kittelsen, Frode Lindseth
19/2009	**Closing the Gates? Evidence from a Natural Experiment on Physicians' Sickness Certification**	Simen Markussen
20/2009	**The Effectss of Sick-Leaves on Earnings**	Simen Markussen
25/2009	**Labour Supply Response of a Retirement Earnings Test Reform**	Erik Hernæs, Zhiyang Jia

The Frisch Centre

The Ragnar Frisch Centre for Economic Research is an independent research institution founded by the University of Oslo. The Frisch Centre conducts economic research in co-operation with the Department of Economics, University of Oslo. The research projects are mostly financed by the Research Council of Norway, government ministries and international organisations. Most projects are co-operative work involving the Frisch Centre and researchers in other domestic and foreign institutions.

Ragnar Frisch Centre for Economic Research
Gaustadalléen 21
N-0349 Oslo, Norway
T + 47 22 95 88 10
F + 47 22 95 88 25
frisch@frisch.uio.no
www.frisch.uio.no