Report
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

Stiftelsen Frischsenteret for samfunnsøkonomisk forskning
Ragnar Frisch Centre for Economic Research

Report 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

Abstract: Firms in three sectors have been subject to random audits by auditors of the Norwegian Tax Authority. The auditing has been carried out according to a detailed procedure securing that the 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 a clear 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 the firms (size, age, location, type of economic activity, use of external auditors, etc.).

Keywords:

Contact: www.frisch.uio.no

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

ISBN 978-82-7988-088-2
ISSN 1501-9721

Contents

Appendices

Summary of main objectives of the overall project and its various parts

Objectives

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 and Part 1 of the main project have been carried out. A report on Part 2 of the main study will be produced later.

The purpose of the preparation study was to explore, discuss and suggest solutions to various issues of a full scale study, and to develop a design to be implemented.[2] The purpose of the pilot study was to test the design on a limited number of audits before a full scale project eventually would be decided. The purpose of Part 1 of the main study has been to implement the developed design on a number of firms sufficient to draw some conclusions about the occurrence and causes of tax evasion.

One goal of the overall project is to estimate the magnitude of tax evasion. Various methods used in different countries to estimate tax evasion have produced widely diverging results, both within and between countries.[3] A main objective is to develop a design of random audits that will improve on existing methods.

A second goal is to estimate how tax evasion varies between industries and various types of firms. Whereas auditors have some knowledge of tax evasion among firms in industries that are often controlled, there are many industries where knowledge is sparse. The objective is to design an audit strategy that will be suitable in most industries.

A third goal is to estimate the effects of sanctions on tax evasion. Rather little is known about the effects of changes both in the probability of tax evasion being revealed and in the severity of sanctions applied. Estimates of these effects might improve on sanction policy.

A forth goal is to develop a design of control strategies that might produce some information about the efficiency of such strategies

A fifth goal is to develop a system of registration of auditing results that may be used for present and future studies of tax evasion and of evaluation of control strategies.

The various parts of the overall project do not include all these goals, cf. the presentation below of Part 1 of the main study.

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

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

[2] Revelation of tax evasion by random audits – Report on the Preparation Study, The Ragnar Frisch Centre, 26. juni 2005.

[3] Erling Eide: Oversikt over litteratur om svart arbeid og skatteunndragelser, Rapport 6/2000, Frischsenteret, s. 89.

Preparation and pilot studies

An overall methodological problem has been to develop an auditing strategy that can produce a sufficient amount of information by use of a reasonable (and limited) amount of auditing resources. The preparation study outlined the main elements of such a strategy. The pilot study refined and developed further this strategy and tested it on a small scale before a full scale project could be decided.

A main feature of the auditing strategy proposed in the preparation study consists of a particular procedure of how to choose the firms to be controlled. An important objective of the pilot project was to test this procedure, which was based on a preliminary distinction between firms with presumed correct tax reporting and firms with presumed incorrect tax reporting. Although the pilot study did not produce very firm conclusions about the fruitfulness of this procedure, it was suggested that it should be applied in the main project.

Another important part of the pilot project was to work out a detailed auditing procedure in cooperation with experienced tax auditors. It has been considered of paramount importance that all tax auditors follow the same procedure, and that the results are recorded in the same manner. A detailed procedure was successfully developed and tested.

Main Project Part 1

A rather comprehensive summary of Part 1 of the Main Study is herby presented in English, whereas the detailed elements of the study are found in Appendices A-G.

The main elements of the study are presented in Section 1 below. The project was extended by a statistical analysis in addition to what was planned (see Appendix G).

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 developed audit strategy (described in section 2 below). Also, the previous parts of the overall project suggested that available resources were sufficient only to audit firms in a few sectors.

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

Section 3 describes how data were obtained from the three sectors that were chosen for investigation.

Section 4 sketches the statistical tests and estimations that have been carried out. The main findings are presented in fat charaters.

In Section 5 we give some advice as to how future parts of the Main Study should be carried out. In Section 6 the costs of the project is presented, together with a list of the collaborators of the project

In Section 7 we present our own evaluation of the project execution.

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

Data on individual tax payers have been anonomysed by “Skattedirektoratet” before the statistical analyses have been carried out.

1 Main elements of the project

1.1 Main elements of the audit strategy

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 developed audit strategy (described in section 2 below). Also, the previous parts of the overall project suggested that available resources were sufficient only to audit firms in a few sectors. It was decided to study three sectors: Wholesale trade of clothing and sport accessories ("Engroshandel med klær, sports- og fritidsutstyr"), freight traffic by road ("godstraffikk på vei"), and cleaning ("renhold"). 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 Authority.

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

The procedure of choosing firms to be controlled has two main elements. First, it was hypothesized – on the basis of informal knowledge – that firms with certain characteristics are more inclined than other firms to report incorrectly. The population of taxpayers (in the three sectors) was partitioned accordingly into firms with *presumably correct reporting* and firms with *presumably incorrect reporting*. Most of the auditing has been carried out in a sample from the last group in order not to "waste" resources on tax payers that do not evade taxes (or evade only to a minor degree). Some firms that are presumed to report correctly have been audited in order to test the procedure.[4]

The second element of the auditing strategy has been to carry out controls in two steps. The first step has consisted in a not very time consuming formal control. In a second step, a sample of the firms has undergone a more detailed control ("bokettersyn"). A main goal of this two-step procedure has been to investigate to which extent the (cheap) formal controls may reveal the existence of tax evasion.

A detailed procedure of how audits should be carried out has been developed and formalised in a PC-program. This main part of the pilot project has been developed at Oslo fylkesskattekontor. The idea has been that all the auditors, when controlling, should be obliged to follow the same procedure and register their findings in boxes supplied by the PC-program.

Statistical tests have been carried out in order to evaluate the fruitfulness of the auditing strategy.

1.2 Variables and statistical analyses

1.2.1 Main statistical analysis

The variable chosen to represent tax evasion is the auditors' proposal (at step 2) of *changes in net income* (changes in relation to the firms' own income tax return). In our main statistical tests and estimations (see Appendix F) this variable is taken as the dependent (response) variable.

[4] Note of March 22, 2004 by Jack Willy Olsen, Sven Tore Christofferen, Karl Børre Reite og Thorild Henriksen describes details of the procedure.

In some of the analyses we are interested only in whether, at step 2, evasion is revealed or not, whether there is a disclosure. In these analyses the dependent variable is a dual (*disclosure* or *not disclosure*).

The explanatory variables are of several types:

- Firms with presumed *correct reports* and firms with presumed *incorrect reports* (dummy variable)
- *External accountant* (dummy variable)
- Three sectors (hereafter called *wholesale, freight traffic*, and *cleaning*, cf. above))
- Five regions of firms' location (*East*, *South*, *West*, *Middle*, and *North*)
- Seven types of municipalities of firms' location (representing industrial structure)
- Size of firms (No. of *employees*)
- Type of organization of firm (corporation or *independent owner* (*sole proprietorship*))
- *Age* of firm
- *"Technical" evaluation* at step 1 of whether evasion will be revealed at step 2 (score constructed on the basis of the proportion of a firm's routines and books that the auditors find unsatisfactory)
- *Auditor's overall evaluation* (marks) at step 1 of whether evasion will be revealed at step 2
- Elements (numbers) of annual reports (net income, operating revenue, business income, sales liable to VAT, cost of rent of premises)

1.2.2 Additional statistical analysis

On the basis of the results of the main statistical analysis (Appendix F) it was decided to study in more detail the *probability of disclosures* (i.e. the probability that the auditors revealed evasion at step 2) and the *expected changes in net income, given disclosure*, as functions of a set of explanatory factors somewhat different from the one described in section 1.2.1 (see Appendix G).

In particular, the following changes were carried out:

- A dummy variable (with two values) representing the "*centrality*" of a municipality was substituted for the regional variable (hereafter: *centrality*)
- A dummy variable (with two values) representing the proportion of service industries in the municipalities was substituted for the seven types of industrial structure in these municipalities (hereafter: *service intensive*)
- The analysis was based on a combination of data sets from the pilot study (2005 data) and (the present) Part 1 of the Main study (2006 data). Accordingly, a dummy variable representing *year* was included.

All explanatory factors are dummy variables with two values, except for an additional factor combining *centrality* and *service intensive*, which has three.

Whereas the main statistical analysis focused (i.a.) on differences in evasion between sectors and between regions, the additional analysis focused on the possible effects on evasion (i.e. the auditor's correction of net income) of the various hypothesised explanatory variables regardless of sector or region.

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 Audit strategy

Even if experience indicates that thorough audits would reveal at least some irregularities in most firms, experience also indicates that some tax payers are reliable in the sense that tax evasion is insignificant or zero. Random audits among all firms might therefore produce a large number audits where no (or insignificant) tax evasion would be found. As a consequence, estimates of tax evasion might be seriously hampered by the small number of observed tax evading firms. Furthermore, there would probably be some reluctance among auditors to participate in a project of random audits where tax evasion in many cases would be insignificant. Moreover, the economic activity in many registered firms has ended without irregularities, and consequently no tax evasion will be observed. We have therefore tested a procedure using the knowledge of auditors to delimit the population in two groups: those which may be assumed to supply correct tax reports and those which may be assumed to supply incorrect ones. In the former group tax evasion is believed to be absent or insignificant. Such a procedure might of course produce skewed estimates. On the other hand the analyses might produce better estimates both because the sample of firms to be audited will be drawn from a smaller population, and because audits probably would be carried out more conscientiously.

As a point of departure for deciding the partition of firms into the two groups we take the view that it is not firms as such that evade taxes, evasion is carried out by some person(s) within the firm, in particular the owner, the managing director, or the chairman of the board. We call persons with such positions in a firm "*main characters*". Instead of directly assuming that firms are reporting correctly or incorrectly, we first divide the main characters into two groups: those who previously have been involved in firms that have reported incorrectly and those who have not been involved in such firms. If a firm has main characters belonging to the first group, we assume that there is a certain risk that tax evasion is taking place.

A firm having at least one main character that previously has been a main character in a firm that has reported incorrectly is presumed also now to be a firm that reports incorrectly. Firms that do not have such main characters are presumed to produce complete reports (not to produce incorrect reports). We consider this system to be a unique and fruitful manner of obtaining and handling individual data on tax payers. A comprehensive description of the whole system is given in the Appendix A.

Random audits have been carried out among samples from both groups, but focus in terms of the *number* of audits has been among the firms that are presumed to report incorrectly reports. Logit analyses have been carried out in order to test the hypothesis that firms presumed to supply correct reports do not evade tax.

Even if the described procedure considerably reduces the population to be studied by random audits, the population will still be large. In order to obtain the most of limited resources, the auditing has been carried out in two steps, as explained above.

3 Data

Data has been collected for the year 2006.

3.1 Selection of firms to be audited

The theoretical basis for selection of firms is given in Appendix B, and the result of the selection is given in Appendix C. It was planned to choose 299 firms for step 1 audits and 99 for step 2 audits. Because some of these firms were out of business in 2006, or turned out to belong to other sectors than those chosen for study, the data consist of 290 firms at step 1 and 90 at step 2. The data obtained by the audits are defined and described in Appendix D.

In the additional statistical analysis (Appendix G) data collected in the Pilot study for the year 2005 has been combined with the 2006 data. The following two data sets were employed:

- Data set 1 (Alternative 1): Wholesale and cleaning for 2005 and wholesale, cleaning, and freight transport for 2006
- Data set 2 (Alternative 2): Wholesale and cleaning for 2005 and 2006.

Data set 1 consisted in step 1 of 419 observations and in step 2 of 137 observations. The data 2 consisted in step 1 of 299 observations and in step 2 of 103 observations.

3.2 Auditors' gathering of data at steps 1 and 2

3.2.1 Formal control of quality of books, step 1

The auditors have at step 1 evaluated the internal control routines and the formal quality of the firms' books and their internal control routines. The auditors have decided that 159 of the firms had to improve their books and internal control routines ("regnskapspålegg"). (These decisions do not constitute an integral part of our statistical analyses.They are more a "by-product.)

3.2.2 Evaluation of formal quality of bookkeeping and internal control routines

On the basis of the detailed reports of the auditors, we have computed a summary statistic, a *"technical" evaluation* ("MaksAvPoeng", MAV), indicating the quality of internal routines and books. This statistic, a score in the interval (0,1), is used as one of the explanatory variables in estimations of evasion.

In addition, the auditors have carried out an *overall evaluation* of whether they expect a firm to evade tax. The auditors have given the firms "marks" according to the following scale (No. of firms in brackets): Satisfactory (142), partly satisfactory, some minor mistakes (81), not satisfactory, serious mistakes/faults (38), Not satisfactory, very serious mistakes/faults (12). Their *overall evaluation* based on these "marks" ("samlet vurdering", SV) is used as another explanatory variable, in addition to or as an alternative to MAP. The evaluation of the auditors is presented in Appendix E.

3.2.3 Comprehensive control, step 2

At step 2, 83 firms have been controlled. (For various reasons 7 out of the 90 firms selected for step 2 controls could not be audited.) The auditors have proposed changes (increases) in net income or in VAT for 19 of the firms. The increases in net income range from NOK 4000 to NOK 4.195.891, with mean NOK 220.836. Appendix D presents summary statistics of the proposed changes. If increases are not proposed, there is no disclosure. The large range of the increases in net income is probably caused by an extreme value (an outlier). In Appendix G

("Rapport 2") the effects of an outlier on various estimates is analyzed. A summary of this analysis is given i Section 4.2.7 below.

The dependent variable to be employed as an indication of tax evasion is the auditors' proposals at step 2 of *changes in net income*. The auditors have recorded various reasons for their proposals, one of which is mistakes in the books about when (in which reporting period) transactions are in fact carried out. Because such mistakes probably are not made for reasons of tax evasion, they are not included in the dependent variable employed. (The name of the resulting variable is "*sum_aarsak*".)

3.3 Other explanatory variables

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

3.4 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 anonymised. Descriptive statistics based on the data file is given in Appendix D.

4 Statistical tests

Two sets of statistical tests have been carried out. The first one includes tests based only on data from 2006, gathered in the present Part 1 of the Main study. The main aspects of the tests and their results are presented in Section 4.1 below. In an additional set of tests data from the Pilot project have been included. These tests and their results are presented in Section 4.2.

4.1 Main statistical tests[5]

4.1.1 Presumed correct vs presumed incorrect reporting

Logistic analyses have been carried out in order to test the procedure of distinguishing between firms that are presumed to report correctly and those which are presumed to report incorrectly (see Section 2 of Appendix F). This is done through comprehensive audits of a random selection of firms of both types. The distinction between the two groups will be considered fruitful if comprehensive audits of the firms drawn from the group in which firms are presumed to report correctly show little or none evasion, whereas such audits reveal widespread evasion in the group of firms that are presumed to report incorrectly.

Comprehensive audits may or may not reveal tax evasion. The result of an audit is represented by a dummy variable Y (the response variable) which takes the value 1 if the audit reveals evasion (a disclosure) and the value 0 if not.

The explanatory factors are represented by three dummy variables. The first represents the assumption about whether the firm is presumed to report correctly or incorrectly. If the firm is presumed to report correctly, a dummy variable I is given the value 1, and 0 otherwise.

[5] This section is a summary of the main elements of Appendix H.

A second dummy variable B represents the three sectors included in the analysis. (B_1 for *wholesale* trade of clothing and sports accessories ("engroshandel med klær, sports- og fritidsutstyr"), B_2 for freight transport by road ("Godstrafikk på vei"); the third sector, cleaning ("renhold") is characterized by $B_1=B_2=0$).

A third dummy variable represents 5 regions (location of firms), consisting of 3-4 districts ("fylker") each. The number of audits in each region appeared to be too low to produce interesting results, and this dummy is not given further consideration here.
The explanatory variables are represented by the vector

$$x = (I, B_1, B_2)\,.$$

The probability that the comprehensive audit will reveal tax evasion given x is

$$p(x) = P(Y = 1 \mid x)$$

In a logistic regression $p(x)$ is transformed to a logit scale, where "logit" is defined by

$$\text{logit}(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right),$$

which is postulated to be linear in x, i.e.:

$$\text{(1)} \qquad \text{logit}(p(x)) = \beta_0 + \beta_1 I + \beta_2 B_1 + \beta_3 B_2\,.$$

The auditing data (from step 2) used include 83 observations of (Y,x), 48 of which are drawn from the population firms assumed to report incorrectly and 35 from the population of firms assumed to report correctly.

The probability that audits reveal evasion, i.e. $\mathrm{P}(\mathrm{Y} = 1 \mid \mathrm{x})$, is estimated for both groups of firms.

Findings

Remembering that our study is based on data related to sales activities and a few sectors only, the following conclusions are drawn on the basis of the estimates:

- **There is no evidence (p-value 0.88) for differences between firms assumed to report correctly and those who are assumed to report incorrectly in the changes of the probabilities of disclosures (disclosure means that auditors propose changes (increases) in net income or VAT at step 2). This result does not imply that the distinction between the two groups of firms would not be relevant for other activities than sales and for sectors not studied. However, from a statistical point of view, random audits without such grouping of tax payers have some is advantages (they are easier to carry out).**
- **The probabilities of disclosures is significantly lower (p-value 0.003) for freight transport than for the two other sectors.**
- **There is no evidence (p-value 0.84) that the probabilities of disclosuresare different between *wholesale* and cleaning.**

- **Assuming equal probability of disclosures between (i) firms that are presumed to supply correct reports and those which are presumed to supply incorrect ones, and (ii) *wholesale* and cleaning, the estimated probability of disclosures freight transport is equal to 0.059 (standard error 0.040) and that for the two other sectors 0.408 (standard error 0.070).**

4.1.2 The effect of regions on the influence of the assumption of correct reports/incorrect reports

The effects of (the five) regions were estimated by including dummy variables for these districts in relation (1).

Finding

No evidence was found for districts, except SOUTH, to have an influence on the probability of disclosures (see Section 3 of Appendix F). If a firm is located in SOUTH, the probability of disclosures is higher than for firms located in other districts.

A Fischer test was carried out in order to investigate further whether the effect on evasion of the reliability variable might differ among sectors.

Findings

As to the effect on evasion (disclosures) of the reliability variable, no significant difference among sectors was found (see Section 4 of Appendix F).

However, including also the data from the pilot for the sector cleaning, we obtained rather strong evidence that the variable "incorrect reporting" has a positive effect on disclosures in this sector. Such evidence was not found for the other two sectors. These results suggest that the distinction between firms that are presumed to report correctly and those which are presumed to report incorrectly may be useful for some sectors, but not for others.

4.1.3 The importance of the evaluation variables at step 1

The main purpose of producing the evaluation variables at step 1 is to obtain a kind of screening variables to sort out those firms for which tax evasion is most likely to be revealed at step 2.

In order to test the effect of these variables the following equation was used:

$$(2) \qquad \text{logit}(p(x)) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 SOUTH + \beta_4 MAV + \beta_5 SV$$

Compared to equation (1), the two evaluation variables *MAV* (the "technical" evaluation of books etc. at step 1) and *SV* (the auditors' overall evaluation) have been included. Furthermore, the district variable *SOUTH* is included, in accordance with conclusions above. On the other hand, the reliablity variable *I* is excluded in accordance with the conclusion above that this variable seems to have no significant effect. (This conclusion was confirmed by a special test where *I* tentatively was included in (2)).

Findings

The auditors' overall evaluation (*SV*) is found to be a good predictor of "disclosures", whereas *MAV* does not provide much additional information. Two separate analyses, where *MAV* and *SV*, respectively, are removed form (2), indicate that *SV* is a slightly better predictor of "disclosures" than MAV. The reason for this difference might be that the auditors, when visiting the firms at step 1, obtain some information that is not included in *MAV*. However, in order to obtain a consistent auditing strategy, stripped form the auditors' possibly mistaken impression, it was decided that *MAV* and not *SV* should be used in order to choose which firms to control at step 2.

4.1.4 The effect of other covariates on the influence of the assumption of correct reports/incorrect reports and on the probability of disclosure

The relationship between *Y* (disclosure, i.e. revealed evasion) and *I* (presumed correct or incorrect reporting) may be influenced by various characteristics of firms. In order to elucidate such influence a number of regression analyses including various covariates have been carried out (see Section 6 of Appendix F). For some covariates there are missing values. In a first group of regression analyses only covariates without missing values are included. These covariates are

I (*correct reporting/incorrect reporting*)
B1 (*wholesale*)
B2 (*freight transport*)
External accountant
No. of employees
EAST
SOUTH
WEST
MIDDLE
NORTH
Sole proprietorship
Company
Age of firm
Regional industrial structure

- type 1 and 2
- type 3
- type 4
- type 5
- type 6
- type 7

The test strategy has been to test various sub-models including only some of these covariates against a full model containing all the 19 covariates. The number of possible sub-models of the full model is $2^{19} = 524\ 288$, too many to investigate. 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.

In a first sub-model, where the covariates representing regional industrial structure were excluded, it was found that these covariates were not statistically significant. Furthermore, in a test against the full model, it was found that the same covariates did not represent any significant contribution. Consequently, these covariates were excluded from all remaining tested sub-models.

Findings

- **A great number of sub-models have been studied, using various methods of excluding and including covariates. The following model is considered to be the best in order to predict disclosures:**

  (4) $$\log it(p(x)) = \beta_0 + \beta_1 B_2 + \beta_2 MAV + \beta_3 SOUTH$$

  with estimates

  (5) $$\log it(\hat{p}(x)) = -1.6577 - 2.3496 B_2 + 1.6417 MAV + 1.5563 SOUTH$$

  An LR-test against the full model indicates that almost nothing is lost by excluding all the other covariates.

  Separate tests using also all the covariates for which some observations are missing did produce similar results.

- **The estimated probability of disclosures is greater in region South than in other regions.**
- **The estimated probability of disclosures is greater among firms for which the "screening" variable MAV indicates tax evasion than among other firms.**
- **The estimated probability of disclosures is lower for freight transport that for the other two sectors.**
- **The intervals of confidence of the estimated probabilities of disclosures are rather large, which is not surprising given the small number of observations.**
- **Even if most of the covariates are not found to be statistically significant, one may not draw the conclusion that they in fact do not have some influence. The number of observations is limited, and more data could possibly lead to statistically significant relationships.**

4.1.5 The effect of separate risk factors used to distinguish between the firm characteristics "correct reporting" and "incorrect reporting"

Even if the distinction between firms that are presumed to report correctly and those which are presumed to report incorrectly was found not to be very fruitful,[6] some of the risk factors used to

[6] Whereas the distinction was found not to be fruitful in general, it had some effect in the sector cleaning.

separate the firms into the two groups might have some effect. In the pilot project the following risk factors were used and tested:

- "Lacking or incorrect statements" (of VAT, etc)
- Irregularities in tax/VAT payments ("betalingsanmenrkinger eller manglende proveny")
- Missing/incorrect information in registers ("registeropplysninger")
- Sales related to various VAT rates "omsetning innenfor flere satser"

In the pilot project we concluded that only the first two had any effects, and consequently only these two were used in the present study.

Findings

A Fisher exact test has demonstrated that the probability of disclosures is higher when the risk factor "Lacking or incorrect statements" (of VAT, etc) is present than when it is not present. A similar effect was not found for the other risk factors

4.1.6 Theoretical basis for analyses of evasion (i.e. of proposed changes in net income)

The stratification of data has the effect that the number of observations in each stratum is very low, and in some cases zero. Consequently, it is not convenient/possible to use standard design-based analysis. Instead, a so-called model-based approach is chosen. A comprehensive presentation of the approach is given in Section 8 of Appendix F.

4.1.7 The effect of various covariates of on the probability of disclosures in step 1 and on evasion

A similar analysis as the one carried out for disclosures in section 4.1.4, has been carried out for the *amount* of tax evasion (proposed changes in net income), (see Section 9 of Appendix F). The analysis is based on the preferred model obtained in 4.1.6 (Section 8 of Appendix F).

Findings

- **The screening variable MAV, (i.e. the "*technical" evaluation* variable at step 1) does not have any significant effect on the *amount* of evasion in the cases where evasion is *present* (where there is a "disclosure"). This means that MAV may help in deciding whether a firm is evading tax, but does not help in predicting how much.**
- **It is the other way round for the variable "*external accountant*". The presence of an external accountant seems to have no effect on the *probability* that the firm is evading tax, but has a statistically significant effect on the *amount* of evasion in cases of disclosure. An external accountant significantly reduces the amount evaded.**
- **The type of sector does not have a statistically significant effect on the amount evaded in cases where evasion takes place. The analyses in section 5.4 indicated that evasion is more common in freight transport than in the two other sectors, but in cases where evasion takes place the *amount* of evasion does not seem to differ.**
- **There is some tendency that when evasion takes place, the amount evaded is smaller in *EAST* than in other regions.**

- **The other covariates do not have any statistically significant effect on the amount evaded (in cases where evasion takes place).**

4.1.8 Estimation of evasion (i.e. proposed changes in net income)

The analyses in section 4.1.4 and 4.1.8 (Sections 6 and 9 in Appendix F, respectively) lead to the conclusion that only the following covariates have any effect on the amount of evasion:

- *Wholesale*
- *Freight transport*
- Region *EAST*
- Region *SOUTH*
- Region *WEST*
- *External accountant*
- *No. of employees*
- *Centrality* (type) of district (kommunesentralitet)

These covariates, and only these, have any effect (direct or indirect) on the distribution of the amount evaded, on "*disclosures*", and on MAV (the "technical" screening parameter). However, when as in this study, data are used to choose among a great number of possible sub-models one cannot conclude that some other combinations of covariates might give almost equally good results.

Findings
A summary of the main estimates is given in Table 1 (see Section 9 of Appendix F for specification on regions)

Table 1 Aggregated proposed changes in net income (evasion) (Mill NOK)

Sector	Number of firms in strata	Predicted evasion	Estimated standard error of predicted evasion	Standard error of estimated standard error of predicted evasion	Total standard error of prediction
Wholesale	1056	104.22	7.88	46.34	47.01
Freight transport	5898	57.71	6.14	162.12	162.24
Cleaning	1166	81.84	6.37	28.33	29.04

4.1.9 Main conclusions of main statistical analyses

- **The distinction between firms that are presumed to report correctly and those which are presumed to report incorretly appears not to be fruitful in general. The auditors have revealed tax evasion (i.e. proposed changes in net income) of about the same size in both groups. This result indicates that tax evasion is more common than informed guesses might suggest.**

- **However, the distinction between the two groups of firms seems to have some effect in one of the sectors (cleaning), but not in the other two.**
- **Among the risk factors used to distinguish between the two groups of firms there is some evidence that the variable "lacking or incorrect statements" (of VAT etc.) has a significant effect. There is no evidence that the other risk factors have a similar effect. (Larger samples might, however, give different conclusions.)**
- **There is evidence for regions to have an effect on the probability of evasion (higher for SOUTH, including Vestfold, Buskerud, Telemark, and Vest-Agder, than for other regions).**
- **The probability of disclosures is considerably higher (and statistically significant) in the sector of freight transport than in the two other sectors.**
- **Both evaluation variables (at step 1) have a statistically significant effect on the probability that the auditors will propose changes in net income at step 2.**
- **Two explanatory variables appear to be of particular importance. There is strong evidence for the existence of an external accountant to have a clear negative effect on the amount avoided (given disclosure), but not on the probability of disclosures. There is also a strong evidence for the evaluation variable (at step 1) to have a substantial effect on the probability of disclosure, but not on the amount evaded. This variable is therefore useful as a screening variable when deciding which firms to choose for a comprehensive control.**
- **The probability of disclosures is lower for freight transport than for *wholesale* and cleaning. For those firms that evade, however, the amounts evaded do not differ much (although evasion might be slightly higher in *wholesale*.**
- **The amounts evaded seem on average to be slightly lower in region EAST than in the other regions.**
- **In the chosen model of prediction, the size of the firm, measured by the number of employees, does not have any effect on the probability of disclosures or on the amount evaded. One can not from this result conclude that size is not of importance for evasion. The reason is that size is highly correlated with the presence of external accountant, and that the effect of size is included in the estimated effect of the external accountant variable. Similar relations might be relevant also for other covariates. Thus, statistically significant variables might not represent causal relationships.**

4.2 Additional statistical tests

In the additional statistical analysis presented in Appendix G a number of regression analyses has been carried out according to the procedure described in section 6 of Appendix F, and the full model has been compared to several prediction models were explanatory factors that turned out not to be statistically significant in the full model (or barely so) have been excluded. The main conclusions are presented in the following subsections. A purpose of evaluating some prediction models is to single out parsimonious models that predict disclosures and *changes in net income* in a satisfactory manner.

4.2.1 The effect on disclosures at step 2 of various covariates and of information obtained at step 1

Table 2 gives an overview of which explanatory variables that are included in the various estimated models. (See Appendix G, section 3.1) The dependent variable is disclosure (i.e. whether auditors have made a *correction of net income*) at step 2. In the various prediction models explanatory variables that appeared not to be statistically significant (or barely significant) have been excluded. In theses tests data set 1 has been used.

The two evaluation criteria (the *"technical" evaluation* and *auditors' overall evaluation* of the quality of books and routines of the firm) were found not to be statistically significant when both are included as explanatory variables. (Estimated coefficients and p-values are given in Appendix G.) However, both of them became statistically significant (or nearly so) if only one of them were included at a time. The *"technical" evaluation* factor has been included in the prediction models.

Table 2 Explanatory variables included in various models estimating "disclosures"

Explanatory Variable	**Full model**	**Prediction model 1**	**Prediction model 2**	**Prediction model 3**	**Prediction model 4**
Year	X	X	X	X	X
"Technical" evaluation at step 1	X	X	X	X	
Auditor's overall evaluation at step 1	X				
Wholesale	X	X			
Cleaning	X	X			
Wholesale and cleaning merged			X	X	X
Newly registered (last 3 years)	X				
Corporation or independent owner	X				
Zero employees, or more than zero	X				
External accountant	X				
Centrality of municipality	X	X	X		
Service intensive	X	X	X		
Centrality and service intensive combined	X			X	
Constant	X	X	X	X	X

The dummies representing *year*, industrial sectors, and *service intensive* municipalities were found to be statistically significant in the full model. Consequently, these factors were included in prediction model 1.

The estimates obtained for the two sectors *wholesale* and *cleaning* were quite similar, and they were therefore merged into a combined sector in the prediction models 2, 3, and 4.

In prediction model 3 the *centrality of municipality* and its *industrial structure* (service intensive) were combined into a single dummy variable.

Models 3 and 4 are considered to be possible prediction models. None of them were rejected against the full model. According to the information criterion *BIC* model 4 clearly is better than model 3, whereas the model 3 is slightly better according to criterion *AIC*.

The following variables turned out not to be statistically significant and consequently excluded in the prediction models: *Corporation or independent owner*, *zero employees or more than zero*, *external accountant*, and *newly registered*. It is demonstrated, however, that the

variable *corporation or independent owner* has an effect on the distribution of the *technical evaluation* factor, and may thus have an indirect effect on *disclosures.*

4.2.2 Probability of disclosures

The prediction models 3 and 4 have been used to estimate the effect on *disclosures* for the two sectors *wholesale* and *cleaning* taken together. The results shows that the probability of disclosures has increased from 2005 to 2006, a result that may indicate a correction of how audits have been carried at step 1 in the two years. In prediction model 3 the highest probabilities of disclosures are obtained for non-central municipalities where the service sector is predominant. These results prevail for both low and high values of the "technical evaluation" variable at step 2, where the high probabilities are obtained for the high values of this variable.

Regressions similar to those described here have been carried out by use of data set 2 (Appendix G, section 3.2). The results are not very different from those obtained by use of data set 1 (appendix G, section 3.3).

4.2.3 The effect on disclosures of including wrong use of value added rates

"Disclosures" are so far defined as cases where auditors have found reported net income too low. When carrying out the same type of studies as described above, but including cases representing wrong use of value added rates, the results obtained to a large extent are similar to those already described (see Appendix G, section 3.4). One difference is that the variable *newly registered* comes out as a statistically significant determinant of disclosures at step 2. Furthermore, the effect of the variable representing municipalities' centrality and the importance of their service sector vanishes.

4.2.4 The information content of the two evaluation criteria

In various tests the two information criteria, "*technical evaluation*" and *auditor's overall evaluation*, have been employed, each at a time (see Appendix G, section 3.5). In the main statistical tests of the probability of disclosure at step 2 it appeared that the latter criterion contained some additional information. In an additional test when both criteria were used together, none of them appeared to be statistically significant, although both obtained significance when used alone. Moreover, the *auditor's overall evaluation* seems to contain additional information when the "*technical" evaluation* does not indicate evasion.

4.2.5 The effect on *changes in income* of observations at step 1 and of other covariates

Data set 1 (See Appendix G, Section 4.1)

The effect on *changes in income* has been estimated in a full model and in prediction models excluding explanatory variables that in the full model have been found not to be statistically significant, see Table 3. (The procedure is similar to the one described above estimating disclosures.) Several prediction models have been estimated, some of which are included In Table 3.

Table 3 Explanatory variables included in various models estimating changes in net income

Explanatory Variable	**Full model**	**Prediction model 1**	**Prediction model 2**	**Prediction model 3**
Year	X			
"Technical" evaluation at step 1	X	X		
Auditor's overall evaluation at step 1	X	X		
Auditor's evaluation combined with low "technical evaluation				X
Wholesale	X			
Cleaning	X			
Wholesale and cleaning merged				
Newly registered (last 3 years)	X	X		X
Corporation or independent owner	X			
Zero employees, or more than zero	X			
External accountant	X	X		
Centrality of municipality	X			
Industrial structure (services or other)	X			
Centrality and industrial structure combined	X			
Constant	X	X	X	X

Information criteria and likelihood testing do not clearly conclude which of the two prediction models to prefer.

In the third prediction model an alternative evaluation criterion is introduced. The criterion is used to represent cases when the *auditor's total evaluation* indicates evasion, when the *"technical" evaluation* does not. In a comprehensive discussion it is argued that this model is to be preferred.

Data set 2

Data set 2 produces more or less the same conclusions as data set 1. One difference is that when data set 2 is employed, the effect of *external accountant* appears to be statistically significant.

4.2.6 Application of estimated model: estimation of correction of net income

Table 4 presents estimates of expected changes in net income given "disclosure". Estimates are given for the sectors *wholesale* and cleaning merged, and dataset 2 is applied.

Estimates vary according to the following characteristics of the firms:

- *Independent owner* (or corporation)
- *Newly registered firm* (during last 3 years), or not
- *External accountant*, or not

Estimates varies between 105 000 NOK and 482 000 NOK. The highest estimates are obtained for newly registered firms without external accountant. Estimates do not differ between corporations and firms with sole proprietorship.

Table 4 also presents estimated probabilities of disclosures. These estimates depend (in addition to the three factors already mentioned) on the centrality of the municipality and whether service industries dominate. (CM=1 means that the municipality is located near a centre, CM=0

that it is not. SI=1 means that service industries dominate, SI=0 that they do not.) The table shows that the probability of disclosures is lowest in centrally located municipalities with a low proportion of service industries and highest in non central municipalities dominated by service industries. The probability of disclosures in the latter municipalities is remarkably high.

Table 4 Estimates of expected changes in net income given disclosure and of probability of "disclosure"

Independent owner	Newly registered	Eksternal accountant	Expected correction of net income, given disclosure (1000 NOK.)	Probability of disclosure Centrality of municipality (CM) Predominance of service industries (SI)		
				CM=1 SI=0	CM = SI	CM=0 SI=1
Yes	Yes	Yes	219	0.189	0.373	0.604
		No	482	0.189	0.373	0.604
	No	Yes	106	0.189	0.373	0.604
		No	234	0.189	0.373	0.604
No	Yes	Yes	211	0.154	0.328	0.567
		No	465	0.154	0.328	0.567
	No	Yes	105	0.154	0.328	0.567
		No	231	0.154	0.328	0.567

In order to illustrate the uncertainty of these estimates, one-tail 95% lower confidence limits have been calculated. These estimates are roughly equal to half of the expected changes in net income.

4.2.7 The effect of an outlier

One of the newly registered firms without an external accountant may be characterised as an outlier. (The correction of net income was NOK 1.2 mill., many times higher than the average.) In order to study the effect of this outlier two estimates have been carried out:

(i) One calculation where the outlier is not included

(ii) One calculation where the correction of net income is (hypothetically) set equal to NOK 120 000 (i.e. 1/10 of the recorded amount).

The results are given in Table 5

Table 4 Estimates of expected changes in net income given disclosure, outlier correction

Independent owner	Newly registered	External accountant	Expected correction of net income		
			Original data	Outlier excluded	Outlier replaced
Yes	Yes	Yes	219	189	177
		No	482	321	266
	No	Yes	106	114	118
		No	234	193	178
No	Yes	Yes	211	183	172
		No	465	313	259
	No	Yes	105	112	117
		No	231	192	176

Table 5, compared to Table 4, shows that the outlier has a substantial effect on the estimated changes in net income. One might be tempted to exclude such an outlier. The population might, however, contain other similar outliers, and important information might be lost if such observations are excluded.

4.2.8 Main conclusions of the additional statistical analysis

- **The probability of disclosures (i.e. correction of net income) varies considerably – from 0.15 to 0.60 – according to the type of municipality and whether the firm is newly registered and has an external accountant.**
- **The probability of disclosures is estimated to be lowest in centrally located municipalities with a low proportion of service industries and highest in non central municipalities dominated by service industries.**
- **The estimated correction of net income given disclosure is higher for newly registered firms without external accountant than for firms that are not newly registered and that have an external accountant.**

5 Information relevant for future random audit studies

Our tests based on data from 2005 (the pilot study) and 2006 indicate that the selection strategy based on the distinction between firms that are presumed to supply correct reports and those which are presumed to supply incorrect ones is not as fruitful as expected. Only in the sector cleaning there is slight evidence for this distinction to have an effect. The fact that evasion seems to be more common among firms assumed to report correctly than previously assumed, indicates that a simple selection strategy within each sector might be just as efficient as the one employed in this study. Note, however, that so far only a few sectors and firm activities have been studied.

Except for this conclusion, we suggest that future studies should be based mainly on the same audit strategy as the present one.

6 Staff and costs

6.1 Staff

The project leader has been Professor in Economics, Erling Eide, University of Oslo and the Frisch Centre. The other participating researchers are Senior researcher Oddbjørn Raaum, Frisch Centre and førsteamanuensis Harald Goldstein, University of Oslo. Paul Gunnar Larssen and Jack-Willy Olsen at the Tax Authority have been in charge of developing the structure and the details of the auditing. Other collaborators at the Tax Authority: Jan-Erik Skogmo, Tone Tysse, and Anders Berset.

6.2 Costs and resources employed

The time the auditors have used for the various auditing tasks has been recorded. The time used in auditing the 83 firms at stage 2 was 6409 hours, or about 77 hours per firm.

The pilot project has been financed partly by the Norwegian Research Council and partly by Skattedirektoratet. The Norwegian Research Council has covered the participation by researchers at the Ragnar Frisch Centre for Economic Research (NOK 500), whereas Skattedirektoratet has covered auditing (incl. some administration). The auditors have registered the amount of time used for the various types of auditing. The average time used per audit was 19 hours, plus some travelling time, i.e. about 4 days per audit. Assessing daily costs to be NOK 2000, the total cost of the 291 audits amounts to NOK 2.328.000.

7 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. A great effort has been made in order to develop systematic control strategies and a system of registration of auditing results. The statistical testing has been rather demanding. In addition, a number of tests not foreseen when the project was designed have been carried out.

Some estimates of the magnitude of tax evasion in some sectors have been estimated.Because of a rather limited number of observation, 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 (unprecise) 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.

Appendikser

Appendix A omfatter følgende emner:

- Beskrivelse av populasjonen mht avgrensing til inaktive virksomheter og stratifiseringsvariablene fylke og bransje.
- Fordeling etter risikoskåre som ble benyttet i pilotundersøkelsen
- Analyse av resultater fra pilotstudie mht komponenter i risikoskåre
- Forlag til revidert risikoskåre.
- Sammenheng mellom opprinnelig og eventuell ny risikoskåre.

Appendix B redegjør for det statistiske grunnlaget for utvalget som skal trekkes. Appendix C viser utfallet av trekningen fordelt på sektorer, fylker og bedrifter med feilaktig eller korrekt rapportering. Appendix D redegjør for hvilke variable det er innhentet data for av revisorene. Appendikset inneholder også summariske oversikter over de innhentede data. Appendix E inneholder deskriptiv statistikk over resultatene av kontrollene. Appendix F og G inneholder (i) logistiske studier av hvorvidt skillet mellom bedrifter med feilaktig eller korrekt rapportering er fruktbart i forbindelse med seleksjon av bedrifter for kontroll (ii) en rekke teser av hypoteser om hvilke faktorer som påvirker forekomsten av unndragelser, og (iii) estimater av unndragelsenes omfang.

Appendix A

Paul Gunnar Larssen: Utvalgsplan

Random audit – hovedprosjekt, del 1

Hensikten med dette notatet er å beskrive populasjonen som utvalget trekkes fra og drøfte problemstillinger knyttet til stratifisering og grenseverdier mht risikoskåre. Notatet er delt i fem:

- Beskrivelse av populasjonen mht avgrensing til inaktive virksomheter og stratifiseringsvariablene fylke og bransje.
- Fordeling etter risikoskåre som ble benyttet i pilotundersøkelsen
- Analyse av resultater fra pilotstudie mht komponenter i risikoskåre
- Forlag til revidert risikoskåre.
- Sammenheng mellom opprinnelig og eventuell ny risikoskåre.

Populasjon - hovedprosjekt
Utgangspunktet for undersøkelsen er alle virksomheter i ti fylker i følgende bransjer (hovednæring):

- Rengjøring (74.700)
- Godstransport på veg (60.240)
- Agentur- og engroshandel med henholdsvis klær, sko og sports- og fritidsutstyr (51.160, 51.41, 51.42 og 51.477)

Fylkene som deltar i undersøkelsen er Østfold, Akershus, Buskerud, Vestfold, Telemark, Vest-Agder, Rogaland, Hordaland, Møre og Romsdal, Sør-Trøndelag og Nordland.

Populasjonen omfatter i alt 17807 virksomheter. Når store virksomheter og antatt inaktive virksomheter fjernes gjenstår 8209 virksomheter.

Tabell 1

Bransje - tosiffer * Populasjoner Crosstabulation

Count

		Populasjoner			
		Inaktive virksomheter	Små og mellomstore virksomheter	Store virksomheter	Total
Bransje - tosiffer	Engroshandel med klær, sports- og fritidsutstyr mv.	1434	1085	35	2554
	Godstransport på vei	6350	5922	36	12308
	Rengjøring	1730	1202	13	2945
Total		9514	8209	84	17807

Som store virksomheter regnes virksomheter som tilfredsstiller et av følgende kriterier:

- Mer enn 100 millioner i omsetning i 2005

- Fere enn 100 ansatte
- Mer enn 50 millioner i omsetning og flere enn 20 ansatte

For å identifisere antatt inaktive virksomheter ble det beregnet en inaktivitetsindikator og en aktivitetsindikator. Statuskode S (slettet) og N (nektet registrering) i MVA, og et utvalg statusmeldinger i ER som indikerer at virksomheten er opphørt gir verdien 1 på inaktivitetsindikatoren. Andre får verdien 0. Ansatte, omsetning og inntekt gir verdien 1 på aktivitetsindikatoren. I tillegg får virksomheter hvor innehaver har hele eller deler av inntekten fra egen virksomhet og ikke har roller i andre virksomheter, verdien 1. På bakgrunn av disse to ble det beregnet en aktivitetskode A (aktiv), dersom det ikke var utslag på inaktivitetsindikatoren og det samtidig var utslag på aktivitetsindikatoren. Dersom det var utsalg på inaktivitetsindikatoren, eller ikke var utslag på aktivitetsindikatoren ble aktivitetskoden satt til I (inaktiv).

Tabell 2 viser hvilke type aktivitet som registrert på aktive små- og mellomstor virksomheter.

Tabell 2

Kilde aktivitetskode

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Inntekt fra egen virksomhet	77	,9	,9	,9
	Skattbar inntekt (etterskudsspl) eller lønnsutbetalinger i 2005	57	,7	,7	1,6
	Årsterminpliktige med oms i 2005	125	1,5	1,5	3,2
	Omsetning i 2006 eller ansatte	7950	96,8	96,8	100,0
	Total	8209	100,0	100,0	

Virksomheter med presumptivt feilaktig eller korrekt rapportering

På bakgrunn av data om rolleinnehavere i virksomhetene er det beregnet en risikoscore. Hver virksomhet tildeles risikoscoren til den av rolleinnehaverne med høyest score. Risikoscoren er basert på følgende faktorer:

- Manglende eller uriktige oppgaver (0.42)
- Betalingsanmerkninger eller skyldig proveny (0.22)
- Uoverenstemmelse mellom registrering i mva og LEP, eller lønnsutbetalinger iflg. LTO (0.26)
- Omsetning innenfor flere avgiftssatser (0.10) (øker risikoen for feil)

Dersom alle disse faktorene er til stede settes scoren til 1, mens den settes til 0 der ingen er til stede. I pilotstudien ble grenseverdien for virksomheter med presumptivt feilaktig rapportering satt til 0.3, slik at kun er manglende eller uriktige oppgaver som alene gir en risikoscore over grenseverdien. Ellers må flere av faktorene være til stede for at rolleinnehaveren/virksomheten

skal komme over grenseverdien. Tabell 3 viser hvordan aktive små og mellomstore virksomheter fordeler seg på bedrifter med antatt korrekt eller feilaktig rapportering innenfor de utvalgte bransjene. Vi ser her at den høyeste andelen virksomheter med antatt feilaktig rapportering er innenfor engroshandel, mens den laveste er innenfor rengjøring.

Tabell 3

Bransje - tosiffer * Presumptivt korrekt/feilaktig rapportering Crosstabulation

			Presumptivt korrekt/feilaktig rapp		
			Presumptivt Ikorrekt	Presumptivt feilaktig	Total
Bransje - tosiffer	Engroshandel med klær, sports- og fritidsutstyr mv.	Count	472	613	1085
		% within Bransje - tosiffer	43,5%	56,5%	100,0%
	Godstransport på vei	Count	3962	1960	5922
		% within Bransje - tosiffer	66,9%	33,1%	100,0%
	Rengjøring	Count	864	338	1202
		% within Bransje - tosiffer	71,9%	28,1%	100,0%
Total		Count	5298	2911	8209
		% within Bransje - tosiffer	64,5%	35,5%	100,0%

Tabell 4

Bransje - tosiffer * Risikonivå Crosstabulation

			Risikonivå						
			<0,2	0,2-0,3	03,-0,4	0,4-0,5	0,5-0,6	>0,6	Total
Bransje - tosiffer	Engroshandel med klær, sports- og fritidsutstyr mv.	Count	272	200	486	30	64	33	1085
		% within Bransje - tosiffer	25,1%	18,4%	44,8%	2,8%	5,9%	3,0%	100,0%
	Godstransport på vei	Count	3282	680	1422	71	320	147	5922
		% within Bransje - tosiffer	55,4%	11,5%	24,0%	1,2%	5,4%	2,5%	100,0%
	Rengjøring	Count	635	229	190	41	57	50	1202
		% within Bransje - tosiffer	52,8%	19,1%	15,8%	3,4%	4,7%	4,2%	100,0%
Total		Count	4189	1109	2098	142	441	230	8209
		% within Bransje - tosiffer	51,0%	13,5%	25,6%	1,7%	5,4%	2,8%	100,0%

I tabell 4 er risikoskåren inndelt i intervaller, og vi ser her at det særlig er i intervallet 0,3-0,4 at det er en større andel virksomheter innen engroshandel. Over 0,4 er andelen høyest for rengjøringsvirksomheter, men andelen er relativt lik i alle tre bransjer.

Analyse av pilotdata mht risikokomponenter

På neste side er faktorene som utgjør riskoskåren satt inn i en regresjonsmodell som uavhengige variabler, og utfall av bokettersynet satt inn som avhengige variabler.

Komponentene er ikke en direkte omkoding av risikoskåren som er brukt ifm utvalgstrekningen. For å fange opp alle risikofaktorene, så gis en virksomhet verdien 1 dersom én av rolleinnehaverne i virksomheten har utslag på en av de ”signifikante” komponentene, ikke bare den med høyest risikoskåre.

Til tross for et lite datamateriale (55 bokettersyn, hvorav 7 medførte endringer) ser vi at faktorene manglende eller utriktige oppgaver (Betakoffesienten=0,18) og betalingsanmerknader/skyldig proveny (Betakoffesienten=0,2) har betydelig sterkere utsalgskraft enn de to andre. Sinifikansnivået er dog bare 0.2 og 0.14. For registeravvik (Betakoffesient=0,09), er signifikansnivået så lavt som 0.5. Mht omsetning innenfor flere

avgiftssatser, så er det en negativ sammenheng (Betakoffesient=-0.04) mellom denne faktoren og responsvariablen, og et signifikansnivå på 0.7.

Regresjonsanalyser - pilotdata

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Omsetning innenfor flere satser, Manglende og uriktige oppgaver, Betalingsanmerknader eller manglende proveny, Registeropplysninger[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Endringer dikotomisert

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,285[a]	,081	,008	,33505

a. Predictors: (Constant), Omsetning innenfor flere satser, Manglende og uriktige oppgaver, Betalingsanmerknader eller manglende proveny, Registeropplysninger

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	,496	4	,124	1,105	,365[a]
	Residual	5,613	50	,112		
	Total	6,109	54			

a. Predictors: (Constant), Omsetning innenfor flere satser, Manglende og uriktige oppgaver, Betalingsanmerknader eller manglende proveny, Registeropplysninger

b. Dependent Variable: Endringer dikotomisert

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,052	,085		,615	,542
	Manglende og uriktige oppgaver	,183	,139	,183	1,323	,192
	Betalingsanmerknader eller manglende proveny	,167	,111	,207	1,496	,141
	Registeropplysninger	,060	,115	,089	,521	,605
	Omsetning innenfor flere satser	-,029	,122	-,040	-,235	,815

a. Dependent Variable: Endringer dikotomisert

Alternativ risikoskåre

Disse resultatene kan gi grunn til å endre risikoskåren slik at de faktorene som ga sterkest utslag i pilotdatane gis større vekt, mens faktorene som ikke ga utslag i pliotdataene gis mindre eller ingen vekt. Et alternativ er å konstrukere en risikoskåre som en dikotom variabel der utslag på en av faktorene manglende eller uriktige oppgaver eller betalingsanmerknader/skyldig proveny gir verdien 1, mens andre gis verdien 0.

Den nye risikoskåren er ikke en direkte omkoding av den gamle (selv om dette hadde vært mulig). For å fange opp alle risikofaktorene, så gis en virksomhet verdien 1 dersom én av rolleinnehaverne i virksomheten har utslag på en av de ”signifikante” komponentene, ikke bare den med høyest risikoskåre.

Tabell 5 viser hvordan fordelingen blir innen for de ulike bransjen med en slik inndeling. Vi ser at andelen virksomheter med antatt feilaktig rapportering er lavere, og noe jevnere fordelt etter denne inndelingen enn etter inndelingen i tabell 2. Rangeringen av bransjer, mht andel av virksomheter med feilaktig rapportering, er også den motsatte av den som fremkommer i tabell 2.

Tabell 5:

Bransje - tosiffer * Riskoskåre alt Crosstabulation

			Riskoskåre alt		
			Presumptivt korrekt rapp.	Presumptivt feilaktig rapp.	Total
Bransje - tosiffer	Engroshandel med klær, sports- og fritidsutstyr mv.	Count	933	152	1085
		% within Bransje - tosiffer	86,0%	14,0%	100,0%
	Godstransport på vei	Count	4799	1123	5922
		% within Bransje - tosiffer	81,0%	19,0%	100,0%
	Rengjøring	Count	914	288	1202
		% within Bransje - tosiffer	76,0%	24,0%	100,0%
Total		Count	6646	1563	8209
		% within Bransje - tosiffer	81,0%	19,0%	100,0%

Sammenlinging av gammel og ny risikoskåre

Tabell 6 viser hvordan populasjonen fordeler seg på intervaller innenfor den gamle skåren, og mellom virksomheter med presumptivt fullstendig eller feilaktig rapportering i den alternative risikoskåren. Vi ser at det er i intervallet0.2-0.3 og 0.3-0.4 det er endringer. I intervallet 0.2.-0.3 havner 37.4% i gruppen med virksomheter med presumptivt feilaktig rapportering med den nye risikoskåren. Dette er virksomheter hvor en av rolleinnehaverne har betalingsanmerknader eller skyldig proveny, men som altså havnet under grensa som var satt i pilotprosjektet. Videre ser vi at 84.1% av virksomhetene i nivå 0.3-0.4 havner under risikogrensa i den nye skåren. Dette er de som har hatt en risikoskåre som innebærer kombinasjon av inkonsistente registeropplysninger og omsetning innenfor flere avgiftssatser. Når ikke alle i dette intervallet "flyttes ned" skyldes det at andre av rolleinnehaverne i virksomheten kan ha hatt en risikoskåre som innbefatter faktoren betalingsanmerknader/skyldig proveny.

I tabell 7 ser vi hvordan endringene er innenfor de tre ulike bransjene. 43.8% (475 virksomheter) av virksomhetene innenfor Engroshandel med klær, sports- og fritidsutstyr "flyttes ned" fra 0.3-0.4 til presumptivt korrekt rapportering, mens 1.3% (14 virksomheter) "flyttes opp". Innenfor Godstransport på vei flyttes 19.1% (1134 virskomheter) ned, mens 43.5% (296 virksomheter) "flyttes opp". Innenfor rengjøring flyttes 129% (155 virksomheter) ned, mens 8.7% (105 virksomheter) flyttes opp.

Tabell 6

Risikonivå * Riskoskåre alt Crosstabulation

			Riskoskåre alt		
			Presumptivt korr. rapp.	Presumptivt feil. rapp.	Total
Risikonivå	<0,2	Count	4188	1	4189
		% within Risikonivå	100,0%	,0%	100,0%
		% of Total	51,0%	,0%	51,0%
	0,2-0,3	Count	694	415	1109
		% within Risikonivå	62,6%	37,4%	100,0%
		% of Total	8,5%	5,1%	13,5%
	03,-0,4	Count	1764	334	2098
		% within Risikonivå	84,1%	15,9%	100,0%
		% of Total	21,5%	4,1%	25,6%
	0,4-0,5	Count	0	142	142
		% within Risikonivå	,0%	100,0%	100,0%
		% of Total	,0%	1,7%	1,7%
	0,5-0,6	Count	0	441	441
		% within Risikonivå	,0%	100,0%	100,0%
		% of Total	,0%	5,4%	5,4%
	>0,6	Count	0	230	230
		% within Risikonivå	,0%	100,0%	100,0%
		% of Total	,0%	2,8%	2,8%
Total		Count	6646	1563	8209
		% within Risikonivå	81,0%	19,0%	100,0%
		% of Total	81,0%	19,0%	100,0%

Tabell 7

Risikonivå * Riskoskåre alt * Bransje - tosiffer Crosstabulation

Bransje - tosiffer				Riskoskåre alt: Presumptivt korr. rapp	Riskoskåre alt: Presumptivt fei.l rapp.	Total
Engroshandel med klær, sports- og fritidsutstyr mv.	Risikonivå	<0,2	Count	272	0	272
			% within Risikonivå	100,0%	,0%	100,0%
			% of Total	25,1%	,0%	25,1%
		0,2-0,3	Count	186	14	200
			% within Risikonivå	93,0%	7,0%	100,0%
			% of Total	17,1%	1,3%	18,4%
		03,-0,4	Count	475	11	486
			% within Risikonivå	97,7%	2,3%	100,0%
			% of Total	43,8%	1,0%	44,8%
		0,4-0,5	Count	0	30	30
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	2,8%	2,8%
		0,5-0,6	Count	0	64	64
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	5,9%	5,9%
		>0,6	Count	0	33	33
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	3,0%	3,0%
	Total		Count	933	152	1085
			% within Risikonivå	86,0%	14,0%	100,0%
			% of Total	86,0%	14,0%	100,0%
Godstransport på vei	Risikonivå	<0,2	Count	3281	1	3282
			% within Risikonivå	100,0%	,0%	100,0%
			% of Total	55,4%	,0%	55,4%
		0,2-0,3	Count	384	296	680
			% within Risikonivå	56,5%	43,5%	100,0%
			% of Total	6,5%	5,0%	11,5%
		03,-0,4	Count	1134	288	1422
			% within Risikonivå	79,7%	20,3%	100,0%
			% of Total	19,1%	4,9%	24,0%
		0,4-0,5	Count	0	71	71
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	1,2%	1,2%
		0,5-0,6	Count	0	320	320
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	5,4%	5,4%
		>0,6	Count	0	147	147
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	2,5%	2,5%
	Total		Count	4799	1123	5922
			% within Risikonivå	81,0%	19,0%	100,0%
			% of Total	81,0%	19,0%	100,0%
Rengjøring	Risikonivå	<0,2	Count	635	0	635
			% within Risikonivå	100,0%	,0%	100,0%
			% of Total	52,8%	,0%	52,8%
		0,2-0,3	Count	124	105	229
			% within Risikonivå	54,1%	45,9%	100,0%
			% of Total	10,3%	8,7%	19,1%
		03,-0,4	Count	155	35	190
			% within Risikonivå	81,6%	18,4%	100,0%
			% of Total	12,9%	2,9%	15,8%
		0,4-0,5	Count	0	41	41
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	3,4%	3,4%
		0,5-0,6	Count	0	57	57
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	4,7%	4,7%
		>0,6	Count	0	50	50
			% within Risikonivå	,0%	100,0%	100,0%
			% of Total	,0%	4,2%	4,2%
	Total		Count	914	288	1202
			% within Risikonivå	76,0%	24,0%	100,0%
			% of Total	76,0%	24,0%	100,0%

Appendix B

Harald Goldstein
12. januar 07

Noen momenter om fordeling av utvalg på strata

Tar utgangspunkt i Paul Gunnar Larsens (PGL) notat av 10. jan. 07.

Totalt utvalg trinn I: 300

Totalt utvalg trinn II 100

Antall strata: 66 (11 fylker, 3 bransjer, 2 risikogrupper)

Trinn I

Oppgaven er å fordele totalutvalget, 300, på strataene. Vi kan redusere antall strata til de 6 (bransje) x (risiko)-strataene, ettersom fordeling av utvalg på fylker er en mer administrativ oppgave (med omtrent like store utvalg totalt pr fylke for å fordele arbeidsmengden likelig, antar jeg).

Oppgaven er altså å fordele 300 på 6 strata, (b, r), der b (bransje) = 1,2,3 og r (risiko) = 1, 2:

Tabell 1. Strata

b	*Bransje*	*r*	*Risikogruppe*
1	Engroshandel med klær	1	Korrekt rapportering
2	Godstransport på vei	2	Feilaktig rapportering
3	Rengjøring		

Stratumstørrelsene, N_{br}, er kjente men avhenger av hvordan risikogruppene er definert, basert på hhv:

- *Risikoskåre 1* – Den opprinnelige definisjonen som ble brukt i pilotundersøkelsen.
- *Risikoskåre 2* – En alternativ definisjon foreslått av PGL.

Tabell 2. Stratumstørrelser

Risikoskåre 1 (opprinnelig)				*Risikoskåre 2 (alternativ)*			
B	*r*=1	*r*=2	*Total*	*b*	*r*=1	*r*=2	*Total*
1	472	613	1085	1	933	152	1085
2	3962	1960	5922	2	4799	1123	5922
3	864	338	1202	3	914	288	1202
Sum	5298	2911	8209	*Sum*	6646	1563	8209

Når det gjelder å bestemme utvalgsstørrelsene, n_{br}, for de 6 strataene, skal vi kort se på 4 metoder eller scenarier:

- *Scenarium* 1: Proporsjonal utvelging.
- *Scenarium* 2: Variansjustert "optimal" utvelging.
- *Scenarium* 3: Konfidensintervall- og variansjustert utvelging (kort: KI-bestemt utvelging).
- *Scenarium* 4: KI-bestemt utvelging med hele trinn-I-utvalget trukket fra (r = 2) – gruppen (feilaktig rapportering) som i pilotundersøkelsen.

Scenariene 1 og 2 er de som vanligvis blir foreslått i litteraturen under navnene "proporsjonale" og "optimale" utvalg h.h.v. Scenarium 3 og 4 er mer skreddersydd til situasjonen.

Scenarium 1 (proporsjonale utvalg):
La N_{br}, n_{br} betegne stratusstørrelse og utvalgsstørrelse h.h.v. Total populasjons- og utvalgsstørrelse er da:

$$N = \sum_{b,r} N_{br} \quad (= 8209) \quad \text{og} \quad n = \sum_{b,r} n_{br} \quad (300)$$

Proporsjonale utvalg fås dermed ved

$$(1) \qquad n_{br} = \frac{N_{br}}{N} n, \qquad b = 1,2,3 \quad r = 1,2$$

Denne planen er spesielt anbefalt hvis

- det er først og fremst totalsummen (evtl. gjennomsnittet) av variablene av interesse, *aggregert over hele populasjonen*, man ønsker å estimere, og samtidig som det er flere variable av interesse,
- eller det er bare en variabel av interesse og ingenting er kjent om stratumvariansene for denne.

Resultatet blir her:

Tabell 3. Utvalgstørrelser ved proporsjonal utvelging

Risikoskåre 1 (opprinnelig)				*Risikoskåre 2 (alternativ)*			
B	*r*=1	*r*=2	*Total*	*b*	*r*=1	*r*=2	*Total*
1	17	22	39	1	34	6	40
2	145	72	217	2	175	41	216
3	32	12	44	3	33	11	44
Sum	194	106	300	*Sum*	242	58	300

Denne løsningen virker ikke spesielt gunstig siden det synes å bli for få observasjoner fra risikogruppe 2 samtidig som bransje synes å dominere for mye.

Scenarium 2 (tekstbok "optimalt" utvalg):
Vi vil foreløpig tenke som om hele utvalget vil bli gjenstand for materiell (trinn 2) kontroll. For stratum (b,r) ($b=1,2,3,\quad r=1,2$), la, som over, N_{br}, n_{br} være stratumstørrelse og utvalgsstørrelse h.h.v. I tillegg innfører vi for enhet *i* i stratum (b,r):

$$Y_{br,i}=\begin{cases}1 & \text{hvis enhet } i \text{ unndrar skatt}\\ 0 & \text{ellers}\end{cases}$$

$p_{br}=P(Y_{br,i}=1)$ ved et rent tilfeldig utvalg fra stratum (b,r). Hvis *z* er en vektor av kovariater, skriver vi $p_{br}(z)=P(Y_{br,i}=1\,|\,z)$.

$X_{br,i}$ = beløp unndratt skatt for enhet *i* [der $(X_{br,i}>0)\Leftrightarrow(Y_{br,i}=1)$].

$T_{br}=\sum_{i=1}^{N_{br}}X_{br,i}$, totalt unndratt beløp i stratum (b,r).

$\sigma_{br}^2=\mathrm{var}(X_{br,i})$, total varians in stratum (b,r). ($\sigma_{br}^2=\dfrac{1}{N_{br}-1}\sum_{i=1}^{N_{br}}\left(X_{br,i}-\bar{X}_{br}\right)^2$)

Formålet med undersøkelsen er primært å anslå p_{br} (eventuelt $p_{br}(z)$) samt T_{br}. Det såkalte "optimale" utvalg minimerer variansen til estimatoren for totalen $T=\sum_{br}T_{br}$, og er gitt ved

(2) $$n_{br}=n\cdot\frac{N_{br}\sigma_{br}}{\sum_{kl}N_{kl}\sigma_{kl}}\quad \text{der}\quad n=300.$$

For å kunne benytte denne formelen, må vi vite noe om σ_{br}, dvs. det er nok å kjenne forholdene $\sigma_{br}/\sigma_{b'r'}$. I praksis benytter man mer eller mindre fornuftige gjetninger på disse forholdene,

vanligvis basert på a priori kunnskap kombinert med informasjon fra pilotundersøkelsen. I vår situasjon kunne vi for eksempel resonnere som følger:

La $\bar{\sigma}^2_{br} = \mathrm{var}(X_{br,i} \mid Y_{br,i} = 1)$ betegne variansen bare blant dem som har unndratt skatt, og $\bar{X}_{U,br} = E(X_{br,i} \mid Y_{br,i} = 1)$ gjennomsnittlig beløp unndratt blant dem som unndrar.
Det kan da vises (se appendiks)

(3) $$\sigma^2_{br} \approx p_{br}\bar{\sigma}^2_{br} + p_{br}(1-p_{br})\bar{X}^2_{U,br} \approx p_{br}\left(\bar{\sigma}^2_{br} + \bar{X}^2_{U,br}\right)$$

Vår (grove) gjetning består nå i å anta

(i) $\bar{\sigma}^2_{br}$ er av samme størrelsesorden i de 6 strataene. Likeledes antas de 6 $\bar{X}^2_{U,br}$ å være av samme størrelsesorden.
(ii) Basert på resultatene fra pilotundersøkelsen (jfr. tabell 3 i appendiks 2 i pilotrapporten), antas (selv om piloten til dels dreier seg om andre bransjer):

$$p_{b2} / p_{b'1} \approx 3 \text{ for } b, b' = 1,2,3$$

Av (i) og (ii) følger nå

(4) $$\sigma^2_{b2}/\sigma^2_{11} \approx 3 \text{ for } b = 1,2,3 \text{ og } \sigma^2_{b1}/\sigma^2_{11} \approx 1 \text{ for } b = 2,3$$
eller

(5) $$\sigma_{b2}/\sigma_{11} \approx 1{,}75 \text{ for } b = 1,2,3 \text{ og } \sigma_{b1}/\sigma_{11} \approx 1 \text{ for } b = 2,3$$

Settes dette inn i (2), får vi

Tabell 4. Utvalgstørrelser ved "optimal" variansjustert utvelging

Risikoskåre 1 (opprinnelig)				*Risikoskåre 2 (alternativ)*			
B	*r*=1	*r*=2	*Total*	*b*	*r*=1	*r*=2	*Total*
1	14	31	45	1	30	9	39
2	114	99	213	2	153	63	216
3	25	17	42	3	29	16	45
Sum	153	147	300	*Sum*	212	88	300

Vi ser at denne løsningen i noen grad imøtegår innvendingen mot løsningen i tabell 3, men fortsatt synes det å være for mye vekt på risikogruppe 1 (korrekt rapportering) og på bransje 2 (godstransport på vei).

Scenarium 3 (KI-bestemt utvalg):

Tabell 3 og 4 bygger på at hovedformålet er estimering av totalen $T = \sum_{br} T_{br}$. Som jeg forstår situasjonen er hovedformålet her heller å estimere stratumtotalene, T_{br} , hver for seg. En alternativ mulighet er derfor å kreve at usikkerheten (uttrykt ved konfidensintervall) skal være den samme i hvert stratum. Nå er usikkerheten (lengden av konfidensintervallet) ved estimering av T_{br} proporsjonal med roten av stratumstørrelsen, N_{br} , som ville føre til uforholdsmessig stor vekt på store strata (i.e. bransje 2), så det synes mer rimelig å kreve at konfidensintervallene for gjennomsnittlig unndragelsesbeløp, $\bar{T}_{br} = T_{br} / N_{br}$, i stedet skal være omtrent like lange. Ved et rent tilfeldig utvalg fra stratum (b,r) blir 95% konfidensintervallet for $\bar{T}_{br}$:

$$\text{estimat}(\bar{T}_{br}) \pm (1,96)\sigma_{br}\sqrt{\frac{N_{br} - n_{br}}{N_{br} n_{br}}}$$

Samme usikkerhet i alle strata innebærer

$$\sigma_{br}^2 \frac{N_{br} - n_{br}}{N_{br} n_{br}} = \sigma_{11}^2 \frac{N_{11} - n_{11}}{N_{11} n_{11}} \quad \text{for} \quad (b,r) \neq (1,1)$$

som gir

$$n_{br} = \frac{N_{br}}{1 + \frac{\sigma_{11}^2}{\sigma_{br}^2} \cdot N_{br} \cdot \frac{N_{11} - n_{11}}{N_{11} n_{11}}}$$

Bruker vi samme gjetning som i (4), får vi dermed

(6) $$n_{b1} = \frac{N_{b1}}{1 + N_{b1} \cdot \frac{N_{11} - n_{11}}{N_{11} n_{11}}} \quad \text{for} \quad b = 2,3$$

(7) $$n_{b2} = \frac{N_{b2}}{1 + N_{b2} \cdot \frac{N_{11} - n_{11}}{N_{11} n_{11}} \cdot \frac{1}{3}} \quad \text{for} \quad b = 1,2,3$$

Sammen med $\sum_{br} n_{br} = 300$, gir dette

Tabell 5. Utvalgstørrelser ved KI-bestemt variansjustert utvelging

Risikoskåre 1 (opprinnelig)				*Risikoskåre 2 (alternativ)*			
B	*r*=1	*r*=2	*Total*	*b*	*r*=1	*r*=2	*Total*
1	26	73	99	1	30	58	88
2	27	79	106	2	31	86	117
3	27	66	93	3	30	70	100
Sum	80	218	298	*Sum*	91	214	305

Totalsummene er de som ligger nærmest 300.

Vi ser at vi her får større vekt på risiko-2 gruppen som er i overensstemmelse med vår forventning om å finne flere unndragelsestilfeller i den gruppen (som vel nettopp er poenget med inndeling i risikogrupper og to-trinns utvelging).

Scenarium 4 (KI-bestemt utvalg med hele trinn-I-utvalget trukket fra de med presumptivt feilaktig rapportering):
I pilotstudien ble hele trinn-I-utvalget trukket fra (*r* = 2)-gruppen (presumptivt feilaktig rapportering), ut fra forventningen at denne gruppen inneholder en større konsentrasjon av enheter som unndrar skatt. La N_b, n_b betegne stratumstørrelse og utvalgsstørrelse for bransje *b* blant enhetene med presumptivt feilaktig rapportering. Bruker vi samme kriterium som i tabell 5 (KI-bestemt), får vi dermed

$$(8) \qquad n_b = \frac{N_b}{1 + N_b \cdot \dfrac{N_1 - n_1}{N_1 n_1}} \quad \text{for } b = 2,3, \text{ der } n_1 + n_2 + n_3 = 300$$

Dette gir

Tabell 6. KI-bestemte utvalg med hele trinn-I-utvalget trukket fra gruppen av virksomheter med presumptivt feilaktig rapportering ((r = 2)-gruppen)

	Risikoskåre 1 (opprinnelig)				*Risikoskåre 2 (alternativ)*			
	$b=1$	$b=2$	$b=3$	Total	$b=1$	$b=2$	$b=3$	Total
N_b	613	1960	338	2911	152	1123	288	1563
n_b	100	113	88	301	74	128	96	298

Trinn II
Utvalget på trinn I blir gjenstand for trinn-I-kontroll. La *Z*=1 (eller $Z_{br,i} = 1$ for en gitt enhet) betegne funn av uregelmessigheter ved trinn-1-undersøkelsen. $Z_{br,i} = 0$ betegner at det ikke er gjort noen funn på trinn I.

På trinn II skal et utvalg på 100 trekkes for en trinn-II-kontroll (materiell kontroll eller bokettersyn). Som i piloten bestemmer *Z* nå 9 post-strata (PS_{kb}):

For bransje, $b = 1, 2, 3$ har vi følgende post-strata:

Poststratum	Definisjon
PS_{1b}	Presumptivt korrekt rapportering $(r = 1)$
PS_{2b}	Presumptivt feilaktig rapporting $(r = 2)$ og $(Z = 0)$
PS_{3b}	Presumptivt feilaktig rapportering $(r = 2)$ og $(Z = 1)$

La $PS_k = PS_{k1} \cup PS_{k2} \cup PS_{k3}$ være poststratum *k* slått sammen over bransje. La m_{kb} betegne utvalgsstørrelsen, og $m_k = m_{k1} + m_{k2} + m_{k3}$ utvalgsstørrelsen i PS_k , der totalt $m = \sum_{kb} m_{kb} = 100$

I pilotstudien valgte man $m_{1b} = m_{2b} + m_{3b} \approx 2m_{2b}$, $m_1 = m_2 + m_3 = 2m_2$, og $m_{k1} \approx m_{k2} \approx m_{k3}$. Her ville denne løsningen innebære $m_1 = 50$ og $m_2 = m_3 = 25$ med $m_{1b} \approx 16$ og $m_{2b} \approx m_{3b} \approx 9$.

Utvalget på $m_1 = 50$ fra PS_1 spilte i piloten utelukkende rolle som kontroll-utvalg for å få noe informasjon til å sammenlikne fordelingen av *Y* og *Z* for enhetermed preumptivt korrekt og feilaktig rapportering. Resultatene fra det nye utvalget vil i tillegg til kontroll bidra til estimeringen av totalene T_{b1}. I piloten var $m_1 = 21$ som kun ga èn $(Y = 1)$-respons (i.e. tilnærmet relativ frekvens 0,05). Hvis denne "tendensen" holder seg vil vi kunne forvente ca 2-3 $(Y = 1)$-responser i vårt utvalg fra virksomheter med preumptivt korrekt rapportering. Med bare ca 16 observasjoner fra en enkelt bransje med presumptivt korrekt rapportering vil det da være stor sjanse for null $(Y = 1)$-responser i minst en av bransjene med presumptivt korrekt rapportering. Dette ville tilsi nødvendigheten av noen ad hoc antakelser av typen som ble gjort ovenfor i forbindelse med tabell 4 for å kunne oppnå estimater med noe fornuft i seg. (Det er her relevant å peke på at T_{b1} i seg selv ikke har primær interesse, men snarere den aggregerte $T_b = T_{b1} + T_{b2}$, aggregert over de to risikogruppene.)

Enten vi bestemmer oss for scenario 3 eller 4, vil det, etter min mening, være en fordel å trekke rent tilfeldig fra *hele* stratumet, PS_{1b}, av virksomheter med presumptivt korrekt rapportering (selv om det innebærer en liten sjanse for at en enhet kan opptre både i trinn-I-utvalget (med scenario 3) og i trinn-II-utvalget). I tillegg mener jeg at alle enhetene i dette utvalget bør registreres både med trinn-I-kontroll (i.e. observer $Z_{b1,i}$) og med full trinn-II-kontroll (i.e. observer $Y_{b1,i}$ og $X_{b1,i}$).

Et naturlig forslag for utvalgsfordeling på trinn II ville nå være å sette $m_1 = 50$, $m_2 = m_3 = 25$ (tilsvarende som i piloten), men istedenfor $m_{k1} \approx m_{k2} \approx m_{k3}$ bruke forholdstallene fra tabell 6, dvs

$$m_{1b} = \frac{50}{300} \cdot n_b \quad \text{og} \quad m_{2b} = m_{3b} = \frac{25}{300} \cdot n_b \quad \text{for} \quad b = 1, 2, 3$$

Dette gir følgende forslag:

Tabell 7. Forslag til utvalgsstørrelser for trinn II

		Risikoskåre 1 (opprinnelig)				*Risikoskåre 2 (alternativ)*			
Stratum	Utvalg	$b=1$	$b=2$	$b=3$	Sum	$b=1$	$b=2$	$b=3$	Sum
PS_{1b}	m_{1b}	16	19	15	50	13	21	16	50
PS_{2b}	m_{2b}	8	9	8	25	6	11	8	25
PS_{3b}	m_{3b}	8	9	8	25	6	11	8	25
	Sum	32	37	31	100	25	43	32	100

Scenario 3 eller 4?

Scenario 3:

Fordel: Ved å slå sammen utvalgene for de enhetene med presumptivt korrekt rapportering fra trinn-I og trinn II får vi mer informasjon til å bestemme fordelinger av *Z* innenfor gruppen med presumptivt korrekt rapportering.

Ulempe: Prisen er et mindre utvalg (informasjon) for fordelinger for *Y* og *X*.

Scenario 4:

Fordel: Større utvalg (informasjon) for fordelinger for *Y* og *X*.
Ulempe: Prisen er et mindre utvalg (informasjon) for fordelinger for *Z*.

Appendiks – Utledning av relasjon (3)

Skriv for korthets skyld *k* for stratum (b,r).
La U_k være mengden av enheter i stratum *k* som unndrar skatt (i.e. med $X_{k,i} > 0$).
La M_k være antall enheter i U_k (som innebærer at $P(Y_{k,i} = 1) = \frac{M_k}{N_k}$).
Vi har

$$\sigma_k^2 = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}\left(X_{k,i} - \bar{X}_k\right)^2 \quad \text{og} \quad \bar{\sigma}_k^2 = \frac{1}{M_k - 1}\sum_{i\in U_k}\left(X_{k,i} - \bar{X}_{U,k}\right)^2$$

Siden $X_{k,i} = 0$ for $i \notin U_k$, får vi

$$\bar{X}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} X_{k,i} = \frac{M_k}{N_k}\cdot\frac{1}{M_k}\sum_{i\in U_k} X_{k,i} = p_k \bar{X}_{U,k}$$

og

$$\sigma_k^2 = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}\left(X_{k,i} - \bar{X}_k\right)^2 = \frac{1}{N_k - 1}\left[\sum_{i\in U_k}\left(X_{k,i} - \bar{X}_k\right)^2 + \sum_{i\notin U_k}\bar{X}_k^2\right] =$$

$$= \frac{1}{N_k - 1}\left[\sum_{i\in U_k}\left(X_{k,i} - \bar{X}_{U,k}\right)^2 + M_k\left(\bar{X}_{U,k} - \bar{X}_k\right)^2 + (N_k - M_k)\bar{X}_k^2\right] =$$

$$= \frac{1}{N_k - 1}\left[(M_k - 1)\bar{\sigma}_k^2 + M_k(1 - p_k)^2\bar{X}_{U,k}^2 + (N_k - M_k)p_k^2\bar{X}_{U,k}^2\right] =$$

$$= \frac{N_k}{N_k - 1}\left[\left(p_k - \frac{1}{N_k}\right)\bar{\sigma}_k^2 + p_k(1 - p_k)\bar{X}_{U,k}^2\right]$$

$$\approx p_k\bar{\sigma}_k^2 + p_k(1 - p_k)\bar{X}_{U,k}^2 \approx p_k(\bar{\sigma}_k^2 + \bar{X}_{U,k}^2)$$

siden $p_k(1 - p_k) \approx p_k$ på grunn av relativt liten p_k.

Appendix C

Utvalgsplan – tillegg

Utvalgsplanen omfatter 33 strata (3 sektorer og 11 fylker). Tabell 1a bygger så direkte som mulig på Tabell 5 i Appendix B, dog med den begrensing at det trekkes 27 objekter per fylke. Tabell 1b viser summen av antall utvalgte enheter i de forskjellige sektorer og fylker, fordelt på bedrifter med presumptivt korrekt og feilaktig rapportering.

I Tabell 2 (a og b) er der lagt til ett objekt per stratum, og det er med utgangspunkt i disse trekksansynlighetene utvalget er trukket. Dette er gjort av hensyn til eventuelle frafall. Da kan vi, i alle fall i første omgang, kompensere for dette uten å måtte trekke på nytt. (En tilfeldig valgt observasjon i hvert stratum settes på "reservebenken").

Tabell 1a Initial utvalgsplan, 3 sektorer, 11 fylker

			Presumptivt korrekt rapp.			Presumptivt feilaktig rapp.		
Fylke	**Bransje**	**Strata**						
Østfold	Engroshandel med klær, sports- og fritidsutstyr r	1	125	3	0,02	13	6	0,46
Østfold	Godstransport på vei	2	562	2	0,00	131	8	0,06
Østfold	Regngjøring	3	65	2	0,03	17	6	0,35
Akershus	Engroshandel med klær, sports- og fritidsutstyr r	4	250	2	0,01	40	8	0,20
Akershus	Godstransport på vei	5	1117	1	0,00	262	7	0,03
Akershus	Regngjøring	6	286	2	0,01	84	7	0,08
Buskerud	Engroshandel med klær, sports- og fritidsutstyr r	7	87	2	0,02	18	7	0,39
Buskerud	Godstransport på vei	8	483	3	0,01	114	6	0,05
Buskerud	Regngjøring	9	97	2	0,02	33	7	0,21
Vestfold	Engroshandel med klær, sports- og fritidsutstyr r	10	89	2	0,02	23	7	0,30
Vestfold	Godstransport på vei	11	310	2	0,01	66	8	0,12
Vestfold	Regngjøring	12	74	2	0,03	22	6	0,27
Telemark	Engroshandel med klær, sports- og fritidsutstyr r	13	28	3	0,11	5	2	0,40
Telemark	Godstransport på vei	14	268	5	0,02	44	8	0,18
Telemark	Regngjøring	15	33	6	0,18	9	3	0,33
Vest-Agde	Engroshandel med klær, sports- og fritidsutstyr r	16	52	3	0,06	8	4	0,50
Vest-Agde	Godstransport på vei	17	206	5	0,02	37	8	0,22
Vest-Agde	Regngjøring	18	51	3	0,06	9	4	0,44
Rogaland	Engroshandel med klær, sports- og fritidsutstyr r	19	103	3	0,03	13	6	0,46
Rogaland	Godstransport på vei	20	362	2	0,01	93	8	0,09
Rogaland	Regngjøring	21	75	2	0,03	28	6	0,21
Hordaland	Engroshandel med klær, sports- og fritidsutstyr r	22	98	3	0,03	16	7	0,44
Hordaland	Godstransport på vei	23	566	2	0,00	152	5	0,03
Hordaland	Regngjøring	24	91	3	0,03	38	7	0,18
Møre og F	Engroshandel med klær, sports- og fritidsutstyr r	25	41	4	0,10	5	2	0,40
Møre og F	Godstransport på vei	26	284	4	0,01	64	8	0,13
Møre og F	Regngjøring	27	42	3	0,07	15	6	0,40
Sør-Trønc	Engroshandel med klær, sports- og fritidsutstyr r	28	48	4	0,08	6	2	0,33
Sør-Trønc	Godstransport på vei	29	389	4	0,01	94	9	0,10
Sør-Trønc	Regngjøring	30	58	3	0,05	15	5	0,33
Nordland	Engroshandel med klær, sports- og fritidsutstyr r	31	12	2	0,17	5	3	0,60
Nordland	Godstransport på vei	32	252	2	0,01	66	9	0,14
Nordland	Regngjøring	33	42	3	0,07	18	8	0,44
			6646	94		1563	203	

Tabell 1b. Initial utvalgsplan, sum sektorer og fylker

	Presumptivt korr. rapp.	Presumptivt feilaktig r.	ialt	Presumptivt korr. rapp.	Presumptivt feilaktig r.	i alt
Engroshandel	31	54	**85**	30	58	*88*
Godstransport	32	84	**116**	31	*86*	*117*
Regnskapsføring	31	85	**96**	30	*70*	*100*
I alt	**94**	**203**	**297**	**91**	***214***	***305***
Østfold	7	20	**27**			
Akershus	5	22	**27**			
Buskerud	7	20	**27**			
Vestfold	6	21	**27**			
Telemark	14	13	**27**			
Vest-Agder	11	16	**27**			
Rogaland	7	20	**27**			
Hordaland	8	19	**27**			
Møre og Rom	11	16	**27**			
Sør-Trøndela	11	16	**27**			
Nordland	7	20	**27**			
I alt	**94**	**203**	**297**			

Tabell 2a. Justert utvalgsplan, 3 sektorer, 11 fylker

Fylke	Bransje	Strata	Presumptivt korrekt rapp. N	n	p	Presumptivt feilaktig rapp. N	n
Østfold	Engroshandel med klær, sports- og fritidsutst	1	125	4	0,03	13	7
Østfold	Godstransport på vei	2	562	3	0,01	131	9
Østfold	Regngjøring	3	66	3	0,05	17	7
Akershus	Engroshandel med klær, sports- og fritidsutst	4	250	3	0,01	40	9
Akershus	Godstransport på vei	5	1117	2	0,00	262	6
Akershus	Regngjøring	6	236	3	0,01	84	8
Buskerud	Engroshandel med klær, sports- og fritidsutst	7	87	3	0,03	18	8
Buskerud	Godstransport på vei	8	483	3	0,01	114	7
Buskerud	Regngjøring	9	97	4	0,04	33	8
Vestfold	Engroshandel med klær, sports- og fritidsutst	10	89	3	0,03	23	8
Vestfold	Godstransport på vei	11	310	3	0,01	66	9
Vestfold	Regngjøring	12	74	3	0,04	22	7
Telemark	Engroshandel med klær, sports- og fritidsutst	13	28	4	0,14	5	3
Telemark	Godstransport på vei	14	269	6	0,02	44	9
Telemark	Regngjøring	15	33	7	0,21	9	4
Vest-Agder	Engroshandel med klær, sports- og fritidsutst	16	52	4	0,08	8	5
Vest-Agder	Godstransport på vei	17	206	6	0,03	37	9
Vest-Agder	Regngjøring	18	61	4	0,06	9	5
Rogaland	Engroshandel med klær, sports- og fritidsutst	19	103	4	0,04	13	7
Rogaland	Godstransport på vei	20	362	3	0,01	93	9
Rogaland	Regngjøring	21	75	3	0,04	28	7
Hordaland	Engroshandel med klær, sports- og fritidsutst	22	96	4	0,04	16	8
Hordaland	Godstransport på vei	23	566	3	0,01	152	6
Hordaland	Regngjøring	24	91	4	0,04	38	8
Møre og Rom	Engroshandel med klær, sports- og fritidsutst	25	41	5	0,12	5	3
Møre og Rom	Godstransport på vei	26	284	5	0,02	54	9
Møre og Rom	Regngjøring	27	42	4	0,10	15	7
Sør-Trøndelag	Engroshandel med klær, sports- og fritidsutst	28	48	5	0,10	6	3
Sør-Trøndelag	Godstransport på vei	29	389	5	0,01	94	10
Sør-Trøndelag	Regngjøring	30	58	4	0,07	15	6
Nordland	Engroshandel med klær, sports- og fritidsutst	31	12	3	0,25	5	4
Nordland	Godstransport på vei	32	252	3	0,01	66	10
Nordland	Regngjøring	33	42	4	0,10	18	9
			6646	127		1563	236

Tabell 2b

Sum for sektorer og fylker

	Presumptivt korrekt rapp.	Presumptivt feilaktig r.	Ialt	Presumptivt korrekt rapp.	Presumptivt feilaktig rapp.	I alt
Engroshandel med klær	42	65	**107**	*30*	*58*	*88*
Godstransport på vei	42	95	**137**	*31*	*86*	*117*
Regngjøring	43	76	**119**	*30*	*70*	*100*
I alt	**127**	**236**	**363**	**91**	***214***	***305***
Østfold	10	23	**33**			
Akershus	8	25	**33**			
Buskerud	10	23	**33**			
Vestfold	9	24	**33**			
Telemark	17	16	**33**			
Vest-Agder	14	19	**33**			
Rogaland	10	23	**33**			
Hordaland	11	22	**33**			
Møre og Romsdal	14	19	**33**			
Sør-Trøndelag	14	19	**33**			
Nordland	10	23	**33**			
I alt	**127**	**236**	**363**			

Appendix D

Paul Gunnar Larssen og Jack Willy Olsen

Datamateriale – Random audit

Identifikasjon
Hver virksomhet har fått et løpenummer (ID).

Filtervariabel – trinn to
Variabelen ja_nei.910 sier hvorvidt det har blitt gjennomført materiell kontroll i virksomheten.

Matriell kontroll på salgsområdet er gjennomført

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ja	83	28,5	92,2	92,2
	Nei	7	2,4	7,8	100,0
	Total	90	30,9	100,0	
Missing	System	201	69,1		
Total		291	100,0		

De som har verdien Nei ble trukket ut til materiell kontroll, men kontrollen har av ulike grunner ikke latt seg gjennomføre. De som har Missing ble ikke trukket ut til trinn to-kontroller.

Avviket mellom planlagte utvalg – 299 trinn én- og 99 trinn to-kontroller skyldes bl.a. virksomheter som er blitt tatt ut av utvalget på grunn av feil bransjekode eller nedleggelse, og som av en eller annen grunn ikke er blitt erstattet.

Koding av verdier
Kodingen av variabler fremkommer også av SPSS-filen (Values).
Ja/Nei variable – 0=Nei, 1=Ja
Ja/Nei/Delvis variable – 1=Ja, 2=Nei, 3=Delvis
Aktuell/Ikke aktuell – 1=Aktuell, 2=Ikke aktuell, 3=Vet ikke
Vurderinger – 0=Ikke aktulell, 1=Tilfredstillende, 2=Tilfredstillene – mindre alvorlige mangler, 3=Ikke tilfredstillende – alvorlige mangler, 4=Ikke tilfredstillende – svært alvorlige mangler
Samlet risikoskåre – 0=Presumptivt korrekt rapportering, 1=Presumptivt feilaktig rapportering.

Missing data
Datasettet inneholder mange obeservasjoner med manglende verdier. Generelt, så vil virksomheter som det ikke har blitt gjennomført trinn to-kontroll ha missing på alle variable knyttet til denne. I trinn én-kontrollen, så gjør revisor en vurdering for hver rutine om denne er aktuell for virksomheten eller ikke. Der rutinen er uaktuell, så vil de andre variablene knyttet til rutinen ha missing (unntatt vurderingen av rutinen, hvor 0 betegner at den ikke er aktuell).

Avhengige variable – trinn to

Vasking av responsvariable
Dataene som har blitt registrert inn i systemet har ikke vært helt konsistente. Det har derfor vært nødvendig å vaske dataene, og å konstruere noen nye variable. Eksempler på inkonsistens er mellom forslag til endring i nettoinntekt, forslag til endring i avgift relatert til endring i nettoinntekt og registrering av årsaker til endring i nettoinntekt. På bakgrunn av de ulike årsaksvariablene har jeg konstruert en ny variable – Sum_aarsak, som er summen av årsaker ekskl. feilperiodisering. Grunnen er at beløp knyttet til feilperiodisering, som for et par enheter er svært store, inngår i forslag til endring i nettoinntekt. Det er tvilsomt om feilperiodisering er en bevisst unndragelse. Det anbefales derfor at Sum_aarsak brukes som numerisk responsvariable framfor forslag til endring i nettoinnekt. Der forslag til endring av nettoinntekt mangler verdi, mens det finnes verdi et av årsaksfeltene er denne verdien imputert til endring i nettoinntekt. Inkonsistens mellom Ja/Nei feltet knyttet til endring i nettoinntekt er også rettet opp. Det er videre noen inkonsistensen mellom endring av avgift knyttet til ending i nettoinntekt. Dette kan være riktig. Da er all endringen knyttet til posten andre endringer. Eksempler på dette er at revisor har laget et skjønn på næringsinntekt, uten at det kan påvises unndratt avgift (det forekommer imidlertid også at revisor bare har oppgitt andre årsaker, og samtidig foreslått endring i merverdiavgift). I de tilfellene hvor det ikke er foreslått endring i merverdiavgift, samtidig som det er utholdt omsetning er angitt som årsak til endring i nettoinntekt, så har jeg beregnet et endringsforslag (25% av utholdt omsetning).

Revisor har bare blitt pålagt å registrere endringer for regnskapsåret 2006. I programmet er det imidlertid åpnet for å også registrere endringer for 2005. Siden registreringen for 2005 er mangelfull, så har jeg sett bort fra den her. Unntaket er der det kun er registrert endringer for 2005. Her har trolig ikke 2006-regnskapet vært tilgjengelig, og i slike tilfeller har jeg imputert 2006-tall på bakgrunn av 2005-tallene.

For de virksomhetene hvor det er gjennomført trinn to-kontroll, så er forslag til endring og årsak til endring satt til 0 der revisor ikke har foreslått noe.

Deskriptiv statistikk, responsvariabler - trinn to
Revisor har både sett på endringer i nettoinntekt og avgift relatert til endring i nettoinntekt, og endringer relatert til feil bruk av mva-satser. Da det bare er tre observasjoner med ending knyttet feil bruk av mva-satser, så er det trolig ikke grunnlag for separate analyser av dette. Jeg har imidlertid konstruert en dikotom responsvariabel (endr – Endringer – alle typer) som sier om revisor har foreslått ending eller ikke (uavhengig av hva slags endring).

Variabelen endring – alle typer viser om revisor har foreslått endring i nettoinntekt og/eller merverdiavgift. Der materiell kontroll er gjennomført har variabler Ja eller Nei som verdi. Ellers Missing.

Enringer - alle typer

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Nei	61	21,0	73,5	73,5
	Ja	22	7,6	26,5	100,0
	Total	83	28,5	100,0	
Missing	System	208	71,5		
Total		291	100,0		

Tabellen under viser deskriptiv statistikk for numeriske variable knyttet til endring i nettoinntekt. Merk at registreringssystemet inneholder flere poster for registrering av årsak til endring. Det er imidlertid bare i de som fremkommer i tabellen at revisor har lagt inn beløp på posten. I datasettet har jeg derfor bare beholdt Ja/Nei variabelen knyttet til disse postene.

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
Endring i nettoinntekt – 2006	17	10797,00	660000,00	3109077,00	182886,8824	166808,07768
Forslag til endring av merverdiavgift, relatert til endinge av nettoinntekt (2006)	13	5255,00	165000,00	696083,00	53544,8462	44432,48432
Sum - årsak til endring av nettoinntetks, ekslk feilperiodisering	19	4000,00	1200000,00	4195891,00	220836,3684	276745,94322
Forslag til endring av merverdiavigift, relatert til feil bruk av mva-satser (2006)	3	22000,00	481555,00	694100,00	231366,6667	232481,19463
Konkret påvist uteholdt omsetning (2006)	13	4000,00	488785,00	2478350,00	190642,3077	154054,92395
Beregnet omsetning ut fra priser og foreliggende vareforbruk/timelister/timebestillinger el.(2006)	1	170000,00	170000,00	170000,00	170000,0000	.
Feilperiodisering av omsetning (2006)	2	1885906,00	1885906,00	3771812,00	1885906,0000	,00000
Andre årsaker (2006)	9	4167,00	1200000,00	1547541,00	171949,0000	389168,83574
Valid N (listwise)	0					

Tabellen under viser deskriptiv statistikk knyttet til feil bruk av mva-satser. Da datamaterialet ikke gir grunnlag for selvstendige analyser av denne responsen, så har jeg ikke gjort noe for å innhente informasjon om årsak til endringer der dette mangler.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Forslag til endring av merverdiavigift, relatert til feil bruk av mva-satser (2006)	3	22000,00	481555,00	231366,6667	232481,19463
Avgpl oms ført som oms utenfor loven (2006)	1	762181,00	762181,00	762181,0000	.
Brukt lavere sats på avpl oms en lovverk tilsier (2006)	1	22000,00	22000,00	22000,0000	.
Valid N (listwise)	0				

Variabler – trinn én

Variablene som er innhentet i forbindelse med trinn én-kontrollen vil dels være forklaringsvariabler knyttet til resultater trinn to-kontrollene, og dels avhengige variable hvor det er mulig å se på et noe større datamateriale.

Innledende spørsmål

Innledningsvis i trinn én-kontrollen skal virksomheten svare på følgende:

- Har virksomhet kun en arbeids-/deltaker?
- Har virksomheten ekstern regnskapsfører?
- Har regnskapsfører ansvar for innbetalinger til bokført kasse mot fysisk kasse?
- Har regnskapsfører ansvar for avstemming av innbetalinger til bokført bank mot bank utskrift?
- Har regnskapsfører ansvar for oppfølging av manglende innbetaling på kundefordringer

Dette er avgjørende for hvilke type kontroller som er aktuelle, og hvilke spørsmål som stilles.

Vurdering av internkontrollrutiner og regnskapspålegg

På bakgrunn av revisors vurdering av internkontrollrutiner og regnskapets formelle kvalitet ble det beregnet en sammendragsfaktor – MaksAvPoengsum Virksomhet, med verdi fra 0 til 1. Følgende tabeller viser henholdsvis frekvens for antall kontroller hvor sammendragsfaktoren var større eller mindre enn 0,3, og deskriptiv statistikk for sammendragsfaktoren.

makspoeng_omk

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<0,3	240	82,5	88,2	88,2
	>0,3	32	11,0	11,8	100,0
	Total	272	93,5	100,0	
Missing	System	19	6,5		
Total		291	100,0		

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
MaksAvPoengsumVirksomhet	273	,00	1,00	,1002	,17457
Valid N (listwise)	273				

Revisor har gitt en samlet vurdering av internkontrollrutinen i virksomheten.

Samlet vurdering

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Ikke aktuell	18	6,2	6,2	6,2
	Tilfredstillende	142	48,8	48,8	55,0
	Delvis tilfredstillende - mindre alvorlige mangler	81	27,8	27,8	82,8
	Ikke tilfredstillende - alvorlige mangler	38	13,1	13,1	95,9
	Ikke tllfredstillende - svært alvorlige manger	12	4,1	4,1	100,0
	Total	291	100,0	100,0	

Det er i tillegg gjort en vurdering av følgende ruiner på tilsvarende skala.

- Revisors vurdering - salgsavtaler
- Revisors vurdering - uttak av varer og tjenester
- Revisors vurdering - registrering av ordre
- Revisors vurdering - dokumentasjon av medgått tid
- Revisors vurdering - dokumentasjon av timebestillinger
- Revisors vurdering - dokumentasjon av varelager
- Revisors vurdering - fastsettelse av priser
- Revisors vurdering - fakturering
- Revisors vurdering - rabatter
- Revisors vurdering - bonuser ol
- Revisors vurdering - kreditnotaer

- Revisors vurdering - manglende innbetaling av kundefordringer
- Revisors vurdering - dokumentasjon av kontantomsetning
- Revisors vurdering - bokførsel -avstemming av innbetalinger til bokført kasse mot fysisk kasse
- Revisors vurdering - bokførsel -avstemming av innbefalinger til bokført bank mot bankutskrifter

I forbindelse med hver av disse rutinene har revisor også ilagt regnskapspålegg dersom dokumentasjon av rutinene ikke er i overensstemmelse med regnskapslovens krav. Tabellen på neste side viser hvor mange virksomheter som har fått regnskapspålegg knyttet til 0, 1, 2, osv. rutiner. Variabelen er også omkodet til to nivåer – har regnkapspålegg og har ikke regnskapspålegg.

ant_regn

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	159	54,6	58,2	58,2
	1,00	65	22,3	23,8	82,1
	2,00	22	7,6	8,1	90,1
	3,00	10	3,4	3,7	93,8
	4,00	6	2,1	2,2	96,0
	5,00	4	1,4	1,5	97,4
	6,00	2	,7	,7	98,2
	7,00	1	,3	,4	98,5
	8,00	3	1,0	1,1	99,6
	14,00	1	,3	,4	100,0
	Total	273	93,8	100,0	
Missing	System	18	6,2		
Total		291	100,0		

I tillegg til dette, som kan kalles respons knyttet til hver rutine er det registrert følgende:

- Er rutinen aktuell for virksomheten (ja/nei)?
- Følges ruiner (ja/nei/delvis)?
- Er rutinen skriftlig dokumentert?

Ansvarlig person – risikoskåre

I forbindelse med hver rutine er det registrert hvilke personer som er ansvarlig for at rutinen følges. Dette kan være en rolleinnehaver eller ansatte i virksomheten. På datasettet er det registrert antall personer som er registrert, samt om en eller flere av personene har en risikoskåre som presumptivt korrekt/feilaktig rapportering (beregnet på samme måte som for innehaver/hovedkarakter i virksomhet). Merk at dersom det er ansatte som er ansvarlige for rutinen, så vil det ikke alltid være mulig å registrere risikoskåre. Risikoskåren beregnes på bakgrunn av historikken til virksomheter personen har hatt en rolle i. Det er derfor mange rutiner hvor risikoskåre mangler.

Nøkkeltall – regnskap

Tabellen på neste side viser summarisk statistikk for nøkkeltallene som blir innhentet gjennom trinn ´en-kontrollene. Det kan se ut til at dataene er relativt komplette for 2005. For 2006 forelå i mange tilfeller ikke regnskapet da trinn én-kontrollene ble gjennomført. Vi kan evt. supplere datasettet med 2006-tall der disse nå er tilgjengelige gjennom DVH (dvs for virksomheter som har levert næringsoppgae elektronisk). Tallene for 2004 er også mangelfulle.

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean	Std. Deviation
Bruttofortjeneste - post 0250/0260 (2006)	64	-14230,00	21307279,00	94577877,12	1477779,3300	3159596,36291
Sum driftsinntekter - post 9000/9900 (2006)	92	47817,00	68644111,00	370666054,53	4028978,8536	9299097,29347
Leie av lokaler - post 6300 (2006)	84	,00	731360,00	4491979,00	53475,9405	130130,72427
Driftsresultat - post 9050/9920 (2006)	81	-508813,00	6085315,00	40110183,50	495187,4506	814165,87636
Næringsinntekt - post 9100/9930 (2006)	76	-519922,00	4253577,00	33977695,93	447074,9464	664065,72947
Bruttofortjeneste - post 0250/0260 (2005)	137	-140203,00	20591993,00	235864817,00	1721641,0000	3409946,15594
Sum driftsinntekter - post 9000/9900 (2005)	238	,00	62803080,00	948935216,00	3987122,7563	8518289,65054
Leie av lokaler - post 6300 (2005)	210	-7876,00	1692226,00	15032899,00	71585,2333	185668,61718
Driftsresultat - post 9050/9920 (2005)	238	-1072238,00	4829455,00	74830437,00	314413,6008	579933,22695
Næringsinntekt - post 9100/9930 (2005)	234	-895678,00	5054804,00	65721639,00	280861,7051	548590,27194
Bruttofortjeneste - post 0250/0260 (2004)	21	-55934,00	20173494,00	77987213,00	3713676,8095	5170997,42243
Sum driftsinntekter - post 9000/9900 (2004)	47	,00	65547221,00	269082985,00	5725169,8936	12574766,71046
Leie av lokaler - post 6300 (2004)	39	,00	1704362,00	5968347,00	153034,5385	323340,73421
Driftsresultat - post 9050/9920 (2004)	48	-526480,00	3183771,00	19461000,00	405437,5000	743922,81899
Næringsinntekt - post 9100/9930 (2004)	48	-772437,00	3008865,00	14545128,00	303023,5000	696990,02173
Valid N (listwise)	0					

Virksomheter med presumptivt korrekt/feilaktig rapportering
I forbindelse med utvalgstrekningen ble hver virksomhet karakterisert som virksomheter med presumptivt korrekt eller feilaktig rapportering på bakgrunn av manglende/uriktige oppgaver og skyldig proveny. I tillegg inneholder datamaterialet riskofaktorene "manglende samsvar mellom registeropplysnigner" og omsetning innenfor flere satser. Vi har avdekket en feil i forbindelse med beregning av risikoskåre. Denne skal beregnes på bakgrunn av historikken til alle virksomheter som hovedkarakterer i virksomheten har hatt en rolle i. I realiteten ble den beregnet på bakgrunn av historikken til virksomheter innenfor årets RA-bransjer som han/hun har hatt en rolle i. Tidligere historikk fra andre bedrifter innehaveren har hatt, fanges dermed ikke opp. Tabellen viser sammenhengen mellom bedrifter med presumptivt korrekt rapportering og presumptivt feilaktig rapportering som lå til grunn for utvalgstrekkingen, og nye beregninger vi har gjort. Vi ser her at 7 virksomheter som ble trukket havnet i feil gruppe. Det vi ikke har kontroll på er hvor mange som fikk mindre sannsynlighet for å bli trukket ut fordi de havnet i gruppa med presumptivt korrekt rapportering.

Risikoskåre - trekkgrunnlag * Risikoskår - rekonstruert Crosstabulation

Count

		Risikoskår - rekonstruert		Total
		Presumptivt korrekt rapportering	Presumptivt feilaktig rapportering	Presumptivt korrekt rapportering
Risikoskåre – trekkgrunnlag	Presumptivt korrekt rapportering	94	7	101
	Presumptivt feilaktig rapportering	0	190	190
Total		94	197	291

Øvrige bakgrunnsvariable

Fra DVH er det trukket ut følgende bakgrunnsvariable:

- Enhetstype Kode
- Forretnings-kommune orgnr
- Hovednaering
- Binaering1
- Binaering2
- Statuskode Enhetsregisteret
- Startet I ER Dato
- Virksomhetens alder
- Antall Ansatte Per Dags Dato (14.12.06)
- Sum Avgpl Oms (Post 2) 2005
- Sum Avgpl Oms (Post 2) 2006 (pr 14.12.06, dvs ca. 5 terminer-av 6)
- Fellesregistrert Virksomhet?
- Sum skattbarinntekt 2005, etterskuddspliktige
- Beløp Lønn(111A) 2005 (virksomhetens lønnsutbetalinger)
- Fylke - forretningsadresse
- Bransje – tosiffer

Merk her at virksomhetens alder er beregnet ut fra oppstartsdato i MVA, ikke i ER, slik at det ikke nødvendigvis er samsvar mellom startet i ER og virksomhetens alder (variablen Startet i ER Dato har mange missing). Variablene Sum Skattbar inntekt 2005 og Beløp Lønn 2005 er hentet fra henholdsvis ligningsregister for etterskuddsplige (LEP) og Lønns- og trekkoppgaveregisteret (LTO). I LEP er det i hovedsak bare AS som er registrert, mens det bare er arbeidgivere (ikke for eksempel ENK uten ansatte) som er registrert i LTO-registeret. De virksomhetene som ikke finnes i disse registrene vil ha missing på disse variablene. Dette var variable som ble hentet inn for å vurdere aktivitet/ikke aktivitet ifm utvalgstrekkingen, men det kan være interessante bakgrunnsvariable.

Tidsbruk
Det er registrert tidsbruk på trinn to-kontrollene knyttet til:

- Tidsforbruk - Arbeid med kontroll av innekt og tilkn. mva
- Tidsforbruk - arbeid med å avdekke kontroll av mva, relatert til feil bruk av mva-satser
- Tidsforbruk - arbeid med å avdekke konkret uteholdt omsetning - fremstilling
- Tidsforbruk - Konkret påvist uteholdt omsetning
- Tidsforbruk - Beregnet omsetning ut fra uteholdt kjøp
- Tidsforbruk - Beregnet omsetning ut fra priser og foreliggende vareforbruk/timelister/timebestillinger el.
- Tidsforbruk - Uriktige kreditnotaer
- Tidsforbruk - Uriktige rabatter
- Tidsforbruk - Uriktige bonuser
- Tidsforbruk - Uriktige tap på fordringer
- Tidsforbruk - Feilperiodisering av omsetning
- Tidsforbruk - Andre årsaker
- Tidsforbruk - Avgpl oms ført som oms utenfor loven
- Tidsbruk - Brukt lavere sats på avpl oms en lovverk tilsier

For disse områdene er det registrert antall timer knyttet til henholdsvis innhenting, analyse og fremstilling. I tillegg er det registrert antall timer til reise, evt. utvidet kontroll og annet. Hensikten med å registrere tidsforbruk er at det skal være mulig å se om, og hvor mye tid revisor har brukt på ulike deler av salgsområdet. Det vil her være mulig å analysere om det er sammenheng mellom revisors vurdering av en rutine/evt regnksapspålegg knyttet til rutinen, hvor mye tid som er brukt på dette i trinn to, og om revisor har foreslått endringer.

Appendix E

Paul Gunnar Larssen

Deskriptiv statistikk, trinn én – random audit, 2007

Trinn én-kontroller
Datamaterialet omfatter data fra 280 trinn én-kontroller. Dette er kontroller hvor revisor går igjennom virksomhetens rutiner på salgsområder. Hver rutine blir vurdert etter følgende skala:

0 – Ikke aktuell
1 – Tilfredsstillende
2 – I all hovedsak tilfredsstillende
3 – I hovedsak ikke tilfredsstillende
4 – Ikke tilfredsstillende

I tillegg til å vurder hver enkelt rutine har revisor også gjort en totalvurdering av virksomhetens rutiner etter samme skala.

Revisor vurderer også hvorvidt regnskapsmateriellet rutinene generer er i overensstemmelse med regnskapsloven. Der de ikke er det ilegges det regnskapspålegg. På bakgrunn av revisors vurdering av de enkelte rutiner og eventuelle regnskapspålegg beregnes det en poengsum med verdier mellom 0 og 1 kalt maks av poeng (MAV). Her beregnes det en sum basert på antall ikke tilfredsstillende rutiner og regnskapspålegg, som divideres på antall aktuelle rutiner. Dersom et flertall av rutinene vurderes som ikke tilfredsstillende og/eller det er ilagt regnskapspålegg vil MAV være høyt. Dersom bare en eller to av mange rutiner ikke er tilfredsstillende vil MAV være lavt.

Revisors samlede vurdering og oppsummeringsfaktor (MAV)

Tabell 1 viser fordeling mellom revisors totalvurdering og MAV. For de fleste med over 0,3 var også totalvurderingen ikke tilfredsstillende. Det var imidlertid et betydelig antall med MAV <0,3 som også ble vurdert som å ha ikke tilfredsstillende rutiner. Til trinn 2-kontrollene ble alle med MAV>0,3 trukket ut, samt et utvalg av de med lavere verdi på denne variabelen.

Tabell 1: Samlet vurdering * makspoeng_omk Crosstabulation

Count

		makspoeng_omk		
		<0,3	>0,3	Total
Samlet vurdering	Ikke aktuell	6	1	7
	Tilfredstillende	136	1	137
	Delvis tilfredstillende - mindre alvorlige mangler	72	7	79
	Ikke tilfredstillende - alvorlige mangler	21	16	37
	Ikke tllfredstillende - svært alvorlige manger	5	7	12
Total		240	32	272

Egenskaper ved virksomhet/MAV

De følgende tabellene viser antall virksomheter hvor trinn én-kontrollene ga utslag over og under grenseverdien for ulike typer virksomheter. Vi ser at det er betydelige forskjeller mellom bransjene. Innenfor rengjøring hadde 21,3% MAV>0,3, mens det for Godstransport på vei kun var 5 som var over denne grenseverdien. 15% av virksomhetene uten ansatte lå over grenseverdien, mens andelen for de med fire eller flere ansatte var i underkant av 5%. Andelen er høyest for de med 0-1 millioner i omsetning, og nær dobbelt så høy for nye virksomheter (1 til 3 år siden oppstart) som for eldre virksomheter. Det er ikke vesentlige forskjeller mellom virksomheter med virksomheter med og uten ekstern regnskapsfører (andelen med MAV over grenseverdi er faktisk litt høyere for virksomheter med ekstern regnskapsfører).

Bransje - tosiffer * makspoeng_omk Crosstabulation

			makspoeng_omk		
			<0,3	>0,3	Total
Bransje - tosiffer	Engroshandel med klær, sports- og fritidsutstyr mv.	Count	64	7	71
		% within Bransje - tosiffer	90,1%	9,9%	100,0%
	Godstransport på vei	Count	102	5	107
		% within Bransje - tosiffer	95,3%	4,7%	100,0%
	Rengjøring	Count	74	20	94
		% within Bransje - tosiffer	78,7%	21,3%	100,0%
Total		Count	240	32	272
		% within Bransje - tosiffer	88,2%	11,8%	100,0%

ans_in * makspoeng_omk Crosstabulation

			makspoeng_omk		
			<0,3	>0,3	Total
ans_in	0 ansatte	Count	113	20	133
		% within ans_in	85,0%	15,0%	100,0%
	1-3 ansatte	Count	68	9	77
		% within ans_in	88,3%	11,7%	100,0%
	4+ ansatte	Count	59	3	62
		% within ans_in	95,2%	4,8%	100,0%
Total		Count	240	32	272
		% within ans_in	88,2%	11,8%	100,0%

oms_int05 * makspoeng_omk Crosstabulation

			makspoeng_omk		
			<0,3	>0,3	Total
oms_int05	0 i oms	Count	2	0	2
		% within oms_int05	100,0%	,0%	100,0%
	-1 mill	Count	97	24	121
		% within oms_int05	80,2%	19,8%	100,0%
	1-5 mill	Count	89	4	93
		% within oms_int05	95,7%	4,3%	100,0%
	5-50 mill	Count	36	1	37
		% within oms_int05	97,3%	2,7%	100,0%
	4,00	Count	5	0	5
		% within oms_int05	100,0%	,0%	100,0%
Total		Count	229	29	258
		% within oms_int05	88,8%	11,2%	100,0%

alder_int1 * makspoeng_omk Crosstabulation

			makspoeng_omk		
			<0,3	>0,3	Total
alder_int1	1-3 år	Count	50	13	63
		% within alder_int1	79,4%	20,6%	100,0%
	4+	Count	190	19	209
		% within alder_int1	90,9%	9,1%	100,0%
Total		Count	240	32	272
		% within alder_int1	88,2%	11,8%	100,0%

Har virksomheten ekstern regnskapsfører? * makspoeng_omk Crosstabulation

			makspoeng_omk		
			<0,3	>0,3	Total
Har virksomheten ekstern regnskapsfører?	Nei	Count	52	6	58
		% within Har virksomheten ekstern regnskapsfører?	89,7%	10,3%	100,0%
	Ja	Count	188	26	214
		% within Har virksomheten ekstern regnskapsfører?	87,9%	12,1%	100,0%
Total		Count	240	32	272
		% within Har virksomheten ekstern regnskapsfører?	88,2%	11,8%	100,0%

Egenskaper ved virksomhet/samlet vurdering

Dersom vi i stedet for den beregnede faktoren ser på revisors subjektive vurdering, så er tendensen stort sett den samme, men andel/antall med ikke tilfredsstillende rutiner er noe høyere (kategori 3 og 4 er gruppert sammen i tabellene). Det er fortsatt høyest andel rengjøringsvirksomheter med ikke tilfredsstillende rutiner. Andelen er også høyest for virksomheter uten ansatte og med 0-1 mill i omsetning. Det er en betydelig større andel virksomheter med ikke tilfredsstillende rutiner i relativt nye enn i eldre virksomheter. Mht bruk av ekstern regnskapsfører, så er imidlertid andel med ikke tilfredsstillende rutiner noe høyere for de uten ekstern regnskapsfører enn med. Dette er motsatt av hva vi fant for MAV. Men forskjellen mellom gruppene er liten.

Bransje - tosiffer * Revisors vurdering - omkodet Crosstabulation

			Revisors vurdering - omkodet		
			Tilfredstillende - helt eller delvis	Ikke tilfredstillende - alvorlige eller svært alvorlige mangler	Total
Bransje - tosiffer	Engroshandel med klær, sports- og fritidsutstyr mv.	Count	52	15	67
		% within Bransje - tosiffer	77,6%	22,4%	100,0%
	Godstransport på vei	Count	102	12	114
		% within Bransje - tosiffer	89,5%	10,5%	100,0%
	Rengjøring	Count	69	23	92
		% within Bransje - tosiffer	75,0%	25,0%	100,0%
Total		Count	223	50	273
		% within Bransje - tosiffer	81,7%	18,3%	100,0%

ans_in * Revisors vurdering - omkodet Crosstabulation

			Revisors vurdering - omkodet		
			Tilfredstillende - helt eller delvis	Ikke tilfredstillende - alvorlige eller svært alvorlige mangler	Total
ans_in	0 ansatte	Count	102	30	132
		% within ans_in	77,3%	22,7%	100,0%
	1-3 ansatte	Count	64	14	78
		% within ans_in	82,1%	17,9%	100,0%
	4+ ansatte	Count	57	6	63
		% within ans_in	90,5%	9,5%	100,0%
Total		Count	223	50	273
		% within ans_in	81,7%	18,3%	100,0%

oms_int05 * Revisors vurdering - omkodet Crosstabulation

			Revisors vurdering - omkodet		
			Tilfredstillende - helt eller delvis	Ikke tilfredstillende - alvorlige eller svært alvorlige mangler	Total
oms_int05	0 i oms	Count	2	0	2
		% within oms_int05	100,0%	,0%	100,0%
	-1 mill	Count	90	30	120
		% within oms_int05	75,0%	25,0%	100,0%
	1-5 mill	Count	84	11	95
		% within oms_int05	88,4%	11,6%	100,0%
	5-50 mill	Count	35	4	39
		% within oms_int05	89,7%	10,3%	100,0%
	4,00	Count	4	0	4
		% within oms_int05	100,0%	,0%	100,0%
Total		Count	215	45	260
		% within oms_int05	82,7%	17,3%	100,0%

alder_int1 * Revisors vurdering - omkodet Crosstabulation

			Revisors vurdering - omkodet		
			Tilfredstillende - helt eller delvis	Ikke tilfredstillende - alvorlige eller svært alvorlige mangler	Total
alder_int1	1-3 år	Count	39	24	63
		% within alder_int1	61,9%	38,1%	100,0%
	4+	Count	184	26	210
		% within alder_int1	87,6%	12,4%	100,0%
Total		Count	223	50	273
		% within alder_int1	81,7%	18,3%	100,0%

Har virksomheten ekstern regnskapsfører? * Revisors vurdering - omkodet Crosstabulation

			Revisors vurdering - omkodet		
			Tilfredstillende - helt eller delvis	Ikke tilfredstillende - alvorlige eller svært alvorlige mangler	Total
Har virksomheten ekstern regnskapsfører?	Nei	Count	44	12	56
		% within Har virksomheten ekstern regnskapsfører?	78,6%	21,4%	100,0%
	Ja	Count	179	38	217
		% within Har virksomheten ekstern regnskapsfører?	82,5%	17,5%	100,0%
Total		Count	223	50	273
		% within Har virksomheten ekstern regnskapsfører?	81,7%	18,3%	100,0%

Personlig risiko og rutiner på salgsområdet

Revisor har foretatt en vurdering av følgende rutiner i bedriftene: Salgsavtaler, uttak av varer og tjenester, registrering av ordre, dokumentasjon av medgått tid, dokumentasjon av timebestillinger, dokumentasjon av varelager, fastsettelse av priser, fakturering, rabatter, bonuser og lignende, kreditnotaer, manglende innbetaling av kundefordringer, ligitimasjon av kontantomsetning, bokførsel – kontanter og bokførsel – bank. Revisor har karakterisert rutinene etter følgende firedelte skala: Tilfredsstillende, tilfredsstillende – mindre alvolige mangler, ikke tilfredsstillende – alvorlige mangler, ikke tilfredsstillende – svært alvorlige mangler. De

ansvarlige for rutinene er gruppert etter deres risikoskår (presumptivt korrekt/feilaktig rapportering). I tillegg til revisors vurderinger er det ilagt regnskapspålegg.

Flest regnskapspålegg ble ilagt knyttet til fakturering. 66 (31,6% av de rutinen var aktuell for) fikk regnskapspålegg knyttet til denne rutinen.
Den høyeste andelen med regnskapspålegg er i forbindelse med dokumentasjon av timebestillinger (50%), men denne rutinen er bare aktuell for 16 virksomheter.
Den sterkeste sammenhengen mellom personlig risikoskåre og regnskapspålegg er for dokumentasjon av varelager. 25% av bedrifter med presumptivt feilaktig rapportering har fått regnskapspålegg knyttet for denne rutiner, mot bare 6,5% for bedrifter med presumptivt korrekt rapportering.

Det er også i forbindelse med fakturering at det er flest virksomheter med mangelfulle rutiner (33). Dette gir også den høyeste andelen (15,6%), hvis vi ser bort fra rutiner som bare er aktuelle for ett mindretall av virksomhetene (uttak av varer og tjenester, legitimasjon av kontantomsetning og bokførsel av kontanter er bare aktuelt for 10-20 virksomheter). Den sterkeste sammenhengen mellom personlig risikoskåre og revisors vurdering av rutinene også for dokumentasjon av varelager. 15% av bedrifter med presumptivt feilaktig rapportering har fått regnskapspålegg knyttet for denne rutiner, mot 6,1% for bedrifter med presumptivt korrekt rapportering.

Deskriptiv statistikk, trinn to – random audit, 2007

Datamaterialet fra trinn to-kontrollene omfatter 83 av de 100 kontrollene som ble trukket ut. Kontrollobjektene ble trukket på bakgrunn av trinn én-kontrollene, ved at alle med MAV>0,3 ble trukket ut. I tillegg ble det trukket et utvalg med MAV<0,3.

Resultater fra trinn to-kontrollene er oppsummert i et excel-vedlegg. I arket **Endringer** (Tabell 2) vises antall og andel med endringer, samt samlet og gjennomsnittlig endringstall (for de med endringer). I tillegg er det beregnet standardavvik og maks- og min-verdier.

Mht endringer i nettoinntekt (eksl feilperiodiseringer) og endring i merverdiavgift relatert til endringer i nettoinntekt er tallene også brutt ned på ulike grupper i de følgende arkene. Da det bare var tre observasjoner med forslag til endring av merverdiavgift knyttet til feil bruk av mva-satser, så er disse tallene ikke brutt ned på ulike grupper.

I arket **Risikoskåre** (Tabell 3a-c) er endringene brutt ned på bedrifter med presumptivt korrekt rapportering og bedrifter med presumptivt feilaktig rapportering, og på ulike risikofaktorer. Det er liten forskjell mellom bedrifter med henholdsvis korrekt og feilaktig rapportering. Forskjellen er størst når det gjelder endring i nettoinntekt (ekskl feilperiodisering). Gjennomsnittsnittlig endring er også størst for bedrifter med presumptivt feilaktig rapportering, men er også variasjonen størst mht beløp. Det er en betydelig forskjell mht risikofaktorene manglende/uriktig oppgave. Andelene med ulike typer endringer er 60, 50 og 40 prosent, mot 22, 16 og 12 for de uten manglende/uriktige oppgaver. Det er ingen, eller til dels negative sammenhenger mellom andel med endringer og de andre risikofaktorene.

Arket **Samlet vurdering** (Tabell 4) viser forholdet mellom forslag til endringer og henholdsvis revisors samlede vurdering og sammendragsfaktorene på bakgrunn av vurderingen av enkeltrutinene (MAV). Der det er blitt avdekket svært alvorlige mangler (4) i trinn én, er det foreslått ending i 100% av tilfellene i trinn to. Det er foreslått endring i nettoinntekt og avgift i 70% av tilfellene. Der det er avdekket alvorlige mangler er andelen med endringer 35% (alle typer endringer), 30% (endringer i nettoinntekt) og 18% (endrnig i avgift). For virksomhetene med mindre alvorlige mangler, eller helt tilfredsstillende rutiner, så er andelen med endringer betydelig lavere. Dersom vi ser på MAV, så er andelen av de med MAV>0,3 med endringer 54,42 og 33, mens andelene for de med MAV<0,3 er 15,12 og 9.

Gjennomsnittlig endringsbeløp (endring i nettoinntekt ekskl feilperiodisering) er noe høyere for de med mindre alvorlige mangler. Men 70% av foreslått endingsbeløp er foreslått i virksomheter hvor trinn én-kontrollene avdekket alvorlige eller svært alvorlige mangler. Tilsvarende andel for virksomheter med MAV>0,3 er 68%.

I arket **Egenskaper** (Tabell 5a-c) har vi sett på forskjeller mellom virksomheter med ulike egenskaper/karakteristika.

Vi ser her at andelen med endringer er noe høyere for Enkeltpersonsforetak enn for selskaper (i hovedsak Aksjeselskaper, men også noen ANS, DA og NUF). Andelen endringer er betydelig høyere for relativt nyetablerte virksomheter (1-3 år), og for virksomheter som ikke har ekstern regnskapsfører. Det ser ikke ut til å være en lineær sammenheng mellom størrelse og andel med endringer dersom vi ser på alle typer endringer. Andel med endringer er imidlertid noe avtagende med størrelse dersom vi ser på endring i nettoinntekt og endring i avgift knyttet til endringer i nettoinntekt (den manglende lineære sammenhengen når det gjelder alle typer endringer skyldes at de tre enhetene med endringer i avgift knyttet til feil bruk av mva-satser er relativt store enheter). Andel endringer er høyest for rengjøringsvirksomheter, mens det bare er to virksomheter med forslag til endring innenfor godstransport på vei. Mht sentralitet, så er andelen endringer høyest i kategori 2B (Noe sentrale kommuner, innenfor 2,5 timer til nivå 3-tettsted) dersom vi ser på alle typer endringer. Dersom vi ser på endring i avgift knyttet til endring i nettoinntekt, så er andelen høyest i de mest sentrale kommunene. Innenfor 2A-kommuner (noe sentrale kommuner, ikke innenfor 2,5 timer til nivå 3-tettsted), så er det bare en virksomhet med forslag til endring,

Revisorenes konklusjoner om regnskapspålegg (Tabell 6) og deres vurdering av virksomhetenes rutiner kan oppsummeres som følger:

Det er for det første ingen lineær sammenheng mellom antall regnskapspålegg og andel med endring. Virksomhetene med bare ett regnskapspålegg er det ingen med endringer, mens 20% av de uten regnskapspålegg har endringer. Av virksomheter med mer enn ett regnskapspålegg vil har 46% endringer.

For alle rutiner, så er det en høyere andel av de som har fått regnskapspålegg knyttet til rutinen, som det også foreligger endringsforslag på.

Dersom vi ser på revisors vurdering av rutiner, så har nesten alle som har fått 4 – svært alvorlige mangler på en av rutinene, og så fått endringsforslag. Unntaket er vurderingen av bokførsel – avstemming av innbetaling til bokført bank mot bankutskrift, hvor tre av fire med svært alvorlige mangler på denne rutinen fått endringsforslag. For nesten alle rutiner, så er andelen med endringsforslag høyere for virksomheter med rutiner med alvorlige mangler (3), enn for virksomheter med tilfredsstillende rutiner eller rutiner med mindre alvorlige mangler.

Tabell 2 Endringer

	Antall trinn to	Endringer - alle	Prosent	Endring i nettoinntekt	Prosent	Endring i MVA	Prosent
Sum alle	83	22	26,5	17	20,5	13	15,7

	Antall	Sum	Mean	Std. Deviation	Minimum	Maximum
Endring i nettoinntetkt*	17	4 171 191	245 364	282 985	10 797	1 200 000
Forslag til endring av merverdiavgiftendinge	13	696 083	53 545	44 432	5 255	165 000
Forslag til endring av merverdiavigift, relatert	3	694 100	231 367	232 481	22 000	481 555

*Sum årsak til endring, eksl feilperiodisering

Tabell 3a Risikoskår

		Antall trinn Antall trinn	Endringer alle typer to	Prosent med endring	Endring i netto-intekt	Prosent med endring	Endring i MVA relatert til endring i nettoinntekt	Prosent med endring
Risikoskår - rekonstruert	Korrekt rapp.	35	9	25,7	6	17,1	5	14,3
	Feilatig rapp.	48	13	27,1	11	22,9	8	16,7
Manglende og uriktige oppgaver	Nei	73	16	21,9	12	16,4	9	12,3
	Ja	10	6	60,0	5	50,0	4	40,0
Betalingsanmerknader eller manglende proveny	Nei	38	12	31,6	9	23,7	8	21,1
	Ja	45	10	22,2	8	17,8	5	11,1
Registeropplysninger	Nei	39	12	30,8	11	28,2	9	23,1
	Ja	44	10	22,7	6	13,6	4	9,1
Omsetning innenfor flere satser	Nei	20	4	20,0	4	20,0	2	10,0
	Ja	63	18	28,6	13	20,6	11	17,5
Sum alle		83	22	26,5	17	20,5	13	15,7

Tabell 3b Risikoskår (forts.)

		Sum - årsak til endring av nettoinntetks, ekslk feilperiodisering				
		Sum	Mean	Std. Deviation	Minimum	Maximum
Risikoskår - rekonstruert	Korrekt rapp.	1 020 294	170 049	120 862	14 786	313 552
	Feilaktig rapp.	3 150 897	286 445	339 949	10 797	1 200 000
Manglende og uriktige oppgaver	Nei	2 013 645	167 804	142 302	10 797	488 785
	Ja	2 157 546	431 509	450 862	72 000	1 200 000
Betalingsanmerknader eller manglende proveny	Nei	1 821 694	202 410	140 326	14 786	430 000
	Ja	2 349 497	293 687	394 330	10 797	1 200 000
Registeropplysninger	Nei	3 293 347	299 395	337 822	10 797	1 200 000
	Ja	877 844	146 307	99 059	14 786	245 364
Omsetning innenfor flere satser	Nei	1 542 443	385 611	547 879	10 797	1 200 000
	Ja	2 628 748	202 211	152 176	14 786	488 785
Sum alle		4 171 191	245 364	282 985	10 797	1 200 000

Tabell 3c Risikoskår (forts.)

Forslag til endring av merverdiavgift, relatert til endinge av nettoinntekt (2006) * Samlet vurdering

		Sum	Mean	Std. Deviation	Minimum	Maximum
Risikoskår - rekonstruert	Korrekt rapp.	251 211	50 242	13 825	29 500	61 343
	Feilaktig rapp.	444 872	55 609	57 119	5 255	165 000
Manglende og uriktige oppgaver	Nei	452 354	50 262	32 449	5 255	122 196
	Ja	243 729	60 932	70 598	7 834	165 000
Betalingsanmerknader eller manglende proveny	Nei	458 125	57 266	47 206	7 834	165 000
	Ja	237 958	47 592	44 170	5 255	122 196
Registeropplysninger	Nei	503 806	55 978	53 441	5 255	165 000
	Ja	192 277	48 069	14 945	29 500	61 343
Omsetning innenfor flere satser	Nei	80 690	40 345	4 992	36 815	43 875
	Ja	615 393	55 945	48 223	5 255	165 000
Sum alle	Total	696 083	53 545	44 432	5 255	165 000

Tabell 4 Samlet vurdering

	A	B	C	D	E	F	G	H	I
1									
2			Antall trinn to	Endringer - all	Prosent med	Endring i nettc	Prosnet mec	Enring i merve	Prosent
3	Samlet vurder	Ikke aktuell	1	1	100,0	1	100,0	1	100,0
4		Tilfredstillend	36	6	16,7	4	11,1	2	5,6
5		Delvis tilfreds	22	2	9,1	2	9,1	2	9,1
6		Ikke tilfredstill	17	6	35,3	5	29,4	3	17,6
7		Ikke tllfredstill	7	7	100,0	5	71,4	5	71,4
8	makspoeng_c	<0,3	59	9	15,3	7	11,9	5	8,5
9		>0,3	24	13	54,2	10	41,7	8	33,3
10									
11			Sum - årsak til endring av nettoinntetks, ekslk feilperiodisering						
12			Sum	Mean	Std. Deviation	Minimum	Maximum	Prosent av endringer	
13	Samlet vurder	Ikke aktuell	175 500	175 500	.	175 500	175 500	4,2	
14		Tilfredstillend	302 175	75 544	110 164	10 797	239 906	7,2	
15		Delvis tilfreds	802 337	401 169	123 908	313 552	488 785	19,2	
16		Ikke tilfredstill	1 769 248	353 850	474 788	72 000	1 200 000	42,4	
17		Ikke tllfredstill	1 121 931	224 386	156 526	27 898	430 000	26,9	
18	makspoeng_c	<0,3	1 349 876	192 839	180 903	10 797	488 785	32,4	
19		>0,3	2 821 315	282 132	341 907	27 898	1 200 000	67,6	
20			4 171 191						
21		Forslag til endring av merverdiavgift, relatert til endinge av nettoinntekt (2006) * Samlet vurdering							
22		Forslag til endring av merverdiavgift, relatert til endinge av nettoinntekt (2006)							
23	Samlet vurder	Samlet vurde	Sum	Mean	Std. Deviation	Minimum	Maximum	Prosent av endringer	
24		Ikke aktuell	43 875	43 875	.	43 875	43 875	6,3	
25		Tilfredstillend	88 434	44 217	20 813	29 500	58 934	12,7	
26		Delvis tilfreds	181 130	90 565	44 733	58 934	122 196	26,0	
27		Ikke tilfredstill	87 149	29 050	18 592	7 834	42 500	12,5	
28		Ikke tllfredstill	295 495	59 099	62 455	5 255	165 000	42,5	
29	makspoeng_c	<0,3	330 907	66 181	33 954	29 500	122 196	47,5	
30		>0,3	365 176	45 647	50 401	5 255	165 000	52,5	
31	Sum alle	Total	696 083	53 545	44 432	5 255	165 000	100,0	
32									

Tabell 5a Egenskaper

		Antall trinn 2	Endringer alle typer	Prosent med endring	Endring i netto-inntekt	Prosent med endring i nettoinntekt	Endring i MVA relatert til endring i nettoinntekt	Prosent
enk_andre	ENK	44	14	31,8	12	27,3	9	20,5
	Selskaper	39	8	20,5	5	12,8	4	10,3
alder_int1	1-3 år	20	10	50,0	8	40,0	5	25,0
	4+	63	12	19,0	9	14,3	8	12,7
ans_in	0 ansatte	44	12	27,3	10	22,7	8	18,2
	1-3 ansatte	22	4	18,2	4	18,2	3	13,6
	4+ ansatte	17	6	35,3	3	17,6	2	11,8
oms_int05	0 i oms	1	0	0,0		0,0		0,0
	-1 mill	42	13	31,0	12	28,6	9	21,4
	1-5 mill	26	4	15,4	3	11,5	2	7,7
	5-50 mill	11	3	27,3	1	9,1	1	9,1
Bransje - tosiffer	Engroshandel r	18	7	38,9	4	22,2	3	16,7
	Godstransport	34	2	5,9	1	2,9	1	2,9
	Rengjøring	31	13	41,9	12	38,7	9	29,0
sentralitet	0A	1	0	0,0	1	100,0	1	100,0
	0B	5	2	40,0		0,0		0,0
	1A	1	0	0,0		0,0		0,0
	1B	1	0	0,0		0,0		0,0
	2A	21	9	42,9	6	28,6	3	14,3
	2B	10	1	10,0	1	10,0		0,0
	3A	44	10	22,7	9	20,5	9	20,5
Har virksomheten ekstern regnskapsfører?	Nei	13	6	46,2	6	46,2	5	38,5
	Ja	70	16	22,9	11	15,7	8	11,4
	Total	83	22	26,5	17	20,5	13	15,7

Tabell 5b Egenskaper (forts.)

		Endring i nettoinntekt					
		Sum	Mean	Std. Deviatic	Minimum	Maximum	Prosent av endringer
enk_andre	ENK	3 320 897	276 741	325 867	10 797	1 200 000	79,6
	Selskaper	850 294	170 059	135 127	14 786	313 552	20,4
alder_int1	1-3 år	2 554 561	319 320	381 368	10 797	1 200 000	61,2
	4+	1 616 630	179 626	150 228	14 786	488 785	38,8
ans_in	0 ansatte	3 086 064	308 606	345 813	27 898	1 200 000	74,0
	1-3 ansatte	516 883	129 221	133 180	10 797	299 400	12,4
	4+ ansatte	568 244	189 415	155 651	14 786	313 552	13,6
oms_int05	0 i oms						0,0
	-1 mill	3 136 261	261 355	322 312	10 797	1 200 000	75,2
	1-5 mill	568 244	189 415	155 651	14 786	313 552	13,6
	5-50 mill	36 686	36 686	.	36 686	36 686	0,9
Bransje - tosiffer	Engroshandel med klær, spo	623 152	155 788	86 882	36 686	245 364	14,9
	Godstransport på vei	488 785	488 785	.	488 785	488 785	11,7
	Rengjøring	3 059 254	254 938	325 604	10 797	1 200 000	73,3
sentralitet	0A	245 364	245 364	.	245 364	245 364	5,9
	2A	661 454	110 242	78 251	10 797	175 500	15,9
	2B	1 200 000	1 200 000	.	1 200 000	1 200 000	28,8
	3A	2 064 373	229 375	168 633	27 898	488 785	49,5
Har virksomheten ekstern regnskapsfører?	Nei	2 286 462	381 077	411 438	72 000	1 200 000	54,8
	Ja	1 884 729	171 339	162 511	10 797	488 785	45,2
	Total	4 171 191	245 364	282 985	10 797	1 200 000	100,0

Tabell 5c Egenskaper (forts.)

		Sum	Mean	Std. Deviat	Minimum	Maximum	Prosent av
enk_andre	ENK	487 372	54 152	53 608	5 255	165 000	70,0
	Selskaper	208 711	52 178	15 161	29 500	61 343	30,0
alder_int1	1-3 år	308 178	61 636	61 214	5 255	165 000	44,3
	4+	387 905	48 488	34 165	7 834	122 196	55,7
ans_in	0 ansatte	472 135	59 017	56 460	5 255	165 000	67,8
	1-3 ansatte	106 080	35 360	6 594	29 500	42 500	15,2
	4+ ansatte	117 868	58 934	0	58 934	58 934	16,9
oms_int05	-1 mill	383 715	42 635	34 571	5 255	122 196	55,1
	1-5 mill	117 868	58 934	0	58 934	58 934	16,9
	5-50 mill	29 500	29 500	.	29 500	29 500	4,2
	Engroshandel med klær, s	133 343	44 448	16 011	29 500	61 343	19,2
Bransje - tosiffer	Godstransport på vei	122 196	122 196	.	122 196	122 196	17,6
	Rengjøring	440 544	48 949	47 470	5 255	165 000	63,3
sentralitet	0B	61 343	61 343	.	61 343	61 343	8,8
	2A	116 192	38 731	7 750	29 817	43 875	16,7
	3A	518 548	57 616	53 271	5 255	165 000	74,5
Har virksomheten ekstern regnskaps	Nei	199 006	39 801	21 758	7 834	61 343	28,6
	Ja	497 077	62 135	53 802	5 255	165 000	71,4
	Total	696 083	53 545	44 432	5 255	165 000	100,0

Tabell 6 Regnskapspålegg trinn 2

	A	B	C	D	E
1			Enringer - alle typer		
2			Nei	Ja	Prosent med endring
3	Regnskapspålegg - salgsavtaler	Nei	33	13	28,3
4		Ja	5	5	50,0
5		**Totalt**	**38**	**18**	**32,1**
6	Regnskapspålegg - uttak av varer og tjenester	Nei	2	0	0,0
7		Ja	4	4	50,0
8		**Totalt**	**6**	**4**	**40,0**
9	Regnskapspålegg - registrering av ordre	Nei	29	12	29,3
10		Ja	3	3	50,0
11		**Totalt**	**32**	**15**	**31,9**
12	Regnskapspålegg - dokumentasjon av medgått tid	Nei	25	7	21,9
13		Ja	7	6	46,2
14		**Totalt**	**32**	**13**	**28,9**
15	Regnskapspålegg - dokumentasjon av timebestillinger	Nei	2	2	50,0
16		**Totalt**	**2**	**2**	**50,0**
17	Regnskapspålegg - dokumentasjon av varelager	Nei	14	5	26,3
18		Ja	5	5	50,0
19		**Totalt**	**19**	**10**	**34,5**
20	Regnskapspålegg - fastsettelse av priser	Nei	48	15	23,8
21		Ja	4	3	42,9
22		**Totalt**	**52**	**18**	**25,7**
23	Regnskapspålegg - fakturering	Nei	36	11	23,4
24		Ja	17	10	37,0
25		**Totalt**	**53**	**21**	**28,4**
26	Regnskapspålegg - rabatter	Nei	11	7	38,9
27		Ja	0	1	100,0
28		**Totalt**	**11**	**8**	**42,1**
29	Regnskapspålegg - bonuser ol	Nei	2	0	0,0
30		Ja	0	1	100,0
31		**Totalt**	**2**	**1**	**33,3**
32	Regnskapspålegg - kreditnotaer	Nei	25	8	24,2
33		Ja	3	4	57,1
34		**Totalt**	**28**	**12**	**30,0**
35	Regnskapspålegg - manglende innbataling av kundefordrir	Nei	32	13	28,9
36		Ja	3	4	57,1
37		**Totalt**	**35**	**17**	**32,7**
38	Regnskapspålegg - dokumentasjon av kontantomsetning	Nei	6	0	0,0
39		Ja	1	3	75,0
40		**Totalt**	**7**	**3**	**30,0**
41	Regnskapspålegg - bokførsel -avstemming av innbetalinge	Nei	3	0	0,0
42		Ja	1	3	75,0
43		**Totalt**	**4**	**3**	**42,9**
44	Regnskapspålegg - bokførsel -avstemming av innbefalinge	Nei	52	12	18,8
45		Ja	5	8	61,5
46		**Totalt**	**57**	**20**	**26,0**
47	ant_regn_int	0	37	9	19,6
48		1	10	0	0,0
49		2-3	9	5	35,7
50		4+	5	7	58,3
51		**Totalt**	**61**	**21**	**25,6**
52		2+	14	12	**46,2**
53					

Appendix F

Harald Goldstein
Resultater fra logistisk analyse

Revidert 23. september 2008

- **Rapport 1: Statistisk analyse – data fra 2006**

- **Analyse av endringer av typen "endringer - alle typer" - dvs. av nettointekt og/eller feil bruk av mva-satser.**

Studien omfatter 291 observasjoner fra 2006 (innhentet 2007) av virksomheter fordelt på 11 fylker (Østfold, Akershus, Buskerud, Vestfold, Telemark, Vest-Agder, Rogaland, Hordaland, Møre og Romsdal, Sør-Trøndelag og Nordland). Utvalget er foretatt i to trinn der trinn 2 består av observasjoner basert på materiell kontroll av 83 virksomheter trukket fra trinn-1-utvalget. En oversikt over utvalg og avdekking finnes i tabell 18 i Appendiks 2.

1. Variable

Framstillingen under er relativt knapp når det gjelder nummerisk beskrivelse av variablene siden mer utførlige deskriptive beskrivelser av variablene er utarbeidet av SKD annet sted (jfr. P.G. Larsen, vedlegg G i hovedrapporten).

Responsvariable

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

Med uttrykket "avdekking" eller "avdekking på trinn 2" menes at $Y = 1$ nedenfor.

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

X og Y opptrer i to versjoner (se P.G. Larsens notat "Datamateriale - Random audit " fra desember 2007). Ingen av versjonene omfatter feilperiodiseringer:

- X_1 ("sum-aarsak") omfatter endringer som er knyttet til endringer i nettoinntekt, men ikke endringer knyttet til feil bruk av mva-satser. Y_1 er en tilsvarende indikator (= 1 hvis $X_1 > 0$ og = 0 ellers).

- X_2 ("endringer - alle typer") omfatter det samme som X_1 pluss endringer knyttet til feil bruk av mva-satser, med Y_2 som tilsvarende indikator.

Denne studien vil konsentrere seg om "endringer - alle typer" slik at *X* og *Y* her svarer til X_2 og Y_2 henholdsvis. X_1 og Y_1 vil bli diskutert i en supplerende rapport.

Blant de 83 virksomhetene trukket ut på trinn 2 var det 22 tilfeller av avdekking av typen "endringer - alle typer" og 18 tilfeller av avdekking av typen "sum-aarsak".

Z - dummy for *funn på trinn 1*, definert ved $Z = \begin{cases} 1 & \text{hvis } \textit{vurderingsskåre} > 0,3 \\ 0 & \text{hvis } \textit{vurderingsskåre} \le 0,3 \end{cases}$

der *vurderingsskåre* er en skåre på skala fra 0 til 1, beregnet på grunnlag av revisors vurdering på trinn 1 av en rekke forhold. Kalt "MaxAvPoeng" i utskrifter.

Variablene *Z* og *Y* vil både opptre som responsvariable og som forklaringsvariable i analysen.

Forklaringsvariable

Presumptivt feilaktig rapportering: $I = \begin{cases} 1 & \text{hvis presumptivt feilaktig rapportering} \\ 0 & \text{ellers} \end{cases}$

Bransje (3 nivåer : 51 -Engroshandel med klær, sports- og fritidsutstyr mv.,
60 - Godstransport på vei,
74 - Rengjøring,
som gir to dummyvariable:
B_1 dummy for engroshandel,
B_2 dummy for godstransport.
Rengjøring er karakterisert ved $B_1 = B_2 = 0$.)

Geografiske forklaringsvariable

Fylke (11 nivåer: Fylke 1,2,6,7,8,10,11,12,15,16,18)

Variabelen *fylke* slås sammen til fem distrikt siden positiv respons ($Y = 1$) forekommer lite i mange av fylkene (0 eller 1 gang i 6 av de 11 fylkene). . Det betyr at det blir for tynt datagrunnlag for en analyse med en individuell dummy for hvert fylke.

Distrikt (Fylke slått sammen til 5 nivåer med dummier):

Øst = dummy for Østfold og Akershus (1 og 2) slått sammen.
Sør = dummy for Buskerud, Vestfold, Telemark og Vest Agder (6, 7, 8 og 10) slått sammen.
Vest = dummy for Rogaland og Hordaland (11 og 12) slått sammen.
Midt = dummy for Møre og Romsdal og Sør-Trøndelag (15 og 16) slått sammen.
Nord = dummy for Nordland.

(Siden disse summerer seg til 1, brukes bare de fire første i en regresjonsmodell.)

Sentralitet av kommunen (SSB) - ordinal skala på 7 nivåer (1 = minst sentral, 7 = mest sentral). Kalt "komsentral" i utskrifter.

Kommunens næringsstruktur (SSB) - nominell skala på 7 nivåer. Beskrives ved dummier kalt "komnæring2", "komnæring3",…., "komnæring7" i utskrifter.

Vurderingsvariable på trinn 1

Vurdering - Samlet vurdering (811) på trinn 1. Ordinal variabel på fire nivåer fra 1 = tilfredstillende til 4 = ikke tilfredstillende (alvorlige feil). Kalt "vurdering_811" i utskrifter..

Vurderingsskåre - En skåre på skala fra 0 til 1, beregnet på grunnlag av revisors vurdering på trinn 1 av en rekke forhold. Kalt "MaxAvPoeng" i utskrifter.

Z - dummy for *funn på trinn 1*, definert ved $Z = \begin{cases} 1 & \text{hvis } \textit{vurderingsskåre} > 0,3 \\ 0 & \text{hvis } \textit{vurderingsskåre} \leq 0,3 \end{cases}$

Andre potensielle forklaringsvariable

Størrelsesmål: antall ansatte pr dags dato, *avgiftspliktig omsetning* (2005) og *driftsinntekter* (2005).
Siden de to siste inneholder en del manglende observasjoner, vil først og fremst *antall ansatte* benyttes som proxy for størrelse (kalt "Ansatte" i utskrifter).

Betydningen av organisasjonsform (ENK og AS definert ved dummier)
Bruk av ekstern regnskapsfører (kalt "EksternRegn" i utskrifter).
Virksomhetens alder - (Kalt "Alder" i utskrifter):

Følgende variable inneholder manglende observasjoner.
Bruttofortjeneste (2005) og (2006)
Sum driftsinntekter (2005) og (2006)
Leie av lokale (2005) og (2006)
Driftsresultat (2005) og (2006)
Næringsinntekt (2005) og (2006)
Avgiftspliktig omsetning (2005) og (2006)

Stratifiserende variable på trinn1

Utvalget på trinn 1 er et stratifisert utvalg trukket fra 66 strata definert av de tre variablene, *presumptivt feilaktig rapportering* (*I*), *bransje* og *fylke*.

2. Vurdering av presumptivt feilaktig rapporterings (I)'s betydning

I modell (1) ser vi først på forklaringsvariablene, $x = (I, B_1, B_2)$. La $p(x) = P("endring" | x) = P(Y = 1 | x)$ betegne sannsynligheten for endring på trinn 2 for en gitt verdi av *x*. I en logistisk regresjon transformeres $p(x)$ til en logit – skala, der "logit" defineres ved

$$\text{logit}(p(x)) = \ln\left(\frac{p(x)}{1 - p(x)}\right)$$

som postuleres å være lineær i *x*, i.e.,

(2.1) $$\text{logit}(p(x)) = \beta_0 + \beta_1 I + \beta_2 B_1 + \beta_3 B_2$$

Merk at modellen ikke postulerer noe om fordelingen til *x*, som er komplisert på grunn av måten utvalget av bedrifter er trukket ut på trinn 1.

Det foreligger *n* = 83 observasjoner av (*Y*,*x*), som fordelte seg i henhold til tabell 2.1

Tabell 2.1 Absolutte hyppigheter

		Endringer - alle typer (Y =1)	Ikke endring (Y = 0)	Sum
Presumptivt korrekt rapportering	Engroshandel Klær, sport- og fritidsutstyr	4	2	6
	Godstransport på vei	1	15	16
	Renhold	4	9	13
Presumptivt feilaktig rapportering	Engroshandel Klær, sport- og fritidsutstyr	3	9	12
	Godstransport på vei	1	17	18
	Renhold	9	9	18
	Sum totalt	22	61	83

Estimering (maximum likelihood) basert på modell (2.1) gir følgende estimater for $p(x) = P(\text{endring} \mid x)$ gitt i tabell 2.2

Tabell 2.2

		Estimert $P(\text{endring}\mid x)$ Modell (2.1)
Presumptivt korrekt rapportering	Engroshandel	0,402
	Godstransport	0,061
	Renhold	0,431
Presumptivt feilaktig rapportering	Engroshandel	0,382
	Godstransport	0,057
	Renhold	0,411

Disse estimatene er basert på regresjonsutskrift 1 i appendiks 2 .

Utskriften og øvrige beregninger viser på grunnlag av modell (2.1):

- Det er ikke evidens (p-verdi 0,88) for forskjell i endringssannsynlighetene mellom bedrifter med presumptiv korrekt rapportering og bedrifter med presumptivt feilaktig rapportering.
- Endringssannsynlighetene er signifikant lavere (p-verdi 0,003) for godstransport enn for de to andre bransjene (estimert oddsfaktor er 0,09 - som sier at oddsen[7] for endring i godstrafikk er ca. 9% av oddsen for endring i de to andre bransjene)..
- Det er ikke evidens (p-verdi 0,84) for at endringssannsynlighetene er forskjellige mellom engroshandel og renhold.
- Antar vi lik endringssannsynlighet mellom bedrifter med presumptivt korrekt rapportering og bedrifter med presumptivt feilaktig rapportering, samt lik endringssannsynlighet i bransjene engroshandel og renhold, blir den estimerte endringssannsynligheten for godstrafikk lik 0,059 (st.avvik 0,040) og for de andre to bransjene lik 0,408 (st.avvik 0,070).

3. Betydning av *distrikt* på sannsynligheten for endring

Observasjonene fordelt på distrikt er gitt i tabell 3.1.

Tabell 3.1

Distrikt	Endring (Y =1)	Ikke endring (Y = 0)	Sum
Øst (Østfold, Akershus)	3	15	18
Sør (Vestfold, Buskerud Telemark, Vest-Agder)	12	18	30
Vest (Rogaland, Hordaland)	2	8	10
Midt (Møre og Romsdal, SørTrøndelag)	4	13	17
Nord (Nordland)	1	7	8
Sum	22	61	83

Vi kan se om *distrikt* har betydning utover *korrekt rapportering/feilaktig rapportering* og *bransje* ved hjelp av modell (3.1)

[7] Merk at hvis *p* betegner sannsynligheten for positiv respons, vil $p/(1-p)$ utgjøre oddsen for det samme.

(3.1) $\text{logit}(p(x)) = \beta_0 + \beta_1 I + \beta_2 B_1 + \beta_3 B_2 + \beta_4 Øst + \beta_5 Sør + \beta_6 Vest + \beta_7 Midt$

som er estimert i utskrift 2 i appendikset.

Likelihood-ratio testing av modell (2.1) (utskrift 1) mot modell (3.1) (utskrift 2) har verdien 2x(40,706 - 37,114) = 7,184, som gir en p-verdi på 0,13 (basert på kji-kvadrat-fordelingen med 4 frihetsgrader). Dvs. det er ikke sterk evidens for at *distrikt* bidrar til forklaring av *Y*. Men det nærmer seg. Koeffisienten for *Sør* er positiv med en p-verdi på 10% som peker i retning av en mulig positiv effekt (høyere sannsynlighet for avdekking) for *Sør*.

I utskrift 3 i appendiks 2 har vi estimert en modell der kun distriktsvariabelen *Sør* er med i tillegg til de andre variablene. Her blir koeffisienten for *Sør* klart signifikant positiv (p-verdi =0,018) som indikerer en høyere sannsynlighet for avdekking i distrikt sør enn de andre distriktene. Koeffisienten for B_2 (godstrafikk) er relativt uberørt av *Sør* og fortsatt sterkt signifikant negativ. Koeffisienten til *I* har nå fått "riktig" fortegn, men er langt fra signifikant forskjellig fra null. Kovariaten *Sør* endrer altså ikke evidensen for at *I* ikke synes å ha betydning for endringssannsynligheten..

4. Varierer effekten av *presumptivt feilaktig rapportering* mellom bransjer?

Modell (1) (og 2) postulerer at effekten av *I* er den samme i alle bransjer. Inspeksjon av tabell 1 kan tyde på at dette kanskje ikke er riktig. I prinsippet kunne vi analysere dette ved å introdusere samspill mellom *I* og bransje i modell 1. Imidlertid ville vi da få en såkalt mettet modell der cellesannsynlighetene i tabell 1 ville falle sammen med de vanlige relative hyppighetene, og siden det er få observasjoner og de estimerte standardavvikene i den logistiske regresjonen er asymptotiske vil disse standardavvikene lett kunne bli misvisende. Her kan og bør vi i stedet bruke Fishers eksakte test som gir eksakte p-verdier uavhengig av sampelstørrelsen.

> [**Fisher's test:** La X_1, X_2 være uavhengige og binomisk fordelte med antall forsøk, n_1, n_2, og suksess-sannsynligheter, p_1, p_2, henholdsvis. Vi skal teste nullhypotesen $H_0: p_1 = p_2$ mot alternativet $H_1: p_1 \neq` p_2$. Testen, som kan vises å ha optimale egenskaper, er en betinget test gitt $Z = X_1 + X_2$, der H_0 forkastes for tilstrekkelig små eller tilstrekkelig store verdier av X_1 i forhold til *Z*. Det viser seg at fordelingen til X_1 (gitt *Z*) under H_0 er helspesifisert hypergeometrisk - som bestemmer p-verdien entydig.]

La oss se på "renhold" først. La indeks 1 stå for *presumptivt feilaktig rapportering* og indeks 2 for *presumptivt korrekt rapportering*. Tabell 1 gir for renhold $X_1 = 9$ og $n_1 = 18$ for bedrifter med presumptivt feilaktig rapportering, og $X_2 = 4$ og $n_2 = 13$ for bedrifter med presumptivt korrekt rapportering. Her blir $Z = X_1 + X_2 = 13$, som

gir p-verdien (eksakt) lik $2 \cdot \min\{P(X_1 \geq 9 \mid Z = 13), P(X_1 \leq 9 \mid Z = 13)\} = 0,48$ hvorav p_1 ikke er signifikant forskjellig fra p_2.

Hvis vi imidlertid antar at endringssannsynlighetene i renhold ikke har endret seg vesentlig fra 2004 til 2006, kan vi kombinere resultatene med resultatene fra pilotundersøkelsen for å få et større materiale. De sammenslåtte dataene[8] ga $(X_1, n_1) = (12, 28)$ for bedrifter med presumptivt feilaktig rapportering og $(X_2, n_2) = (4, 25)$ for bedrifter med presumptivt korrekt rapportering. Dette gir en p-verdi på 0,066 slik at p_1 er signifikant større enn p_2 med signifikansnivå 6,6%. Det er altså rimelig sterk evidens for variabelen *preumptivt feilaktig rapportering* har en positiv effekt i renholdsbransjen.

Når det gjelder "Engroshandel med klær, sports- og fritidsutstyr mv.", ga de foreliggende 2006-dataene $(X_1, n_1) = (3,12)$ for presumptivt feilaktig rapportering og $(X_2, n_2) = (4, 6)$ for presumptivt korrekt rapportering, som leder til p-verdi = 0,24; altså ikke-signifikans. Slår vi sammen med observasjonene fra 2004, blir $(X_1, n_1) = (4,16)$ for presumptivt feilaktig rapportering og $(X_2, n_2) = (5,15)$ for presumptivt korrekt rapportering. Tilsvarende p-verdi blir 0,54 og fortsatt klar ikke-signifikans. Det er altså ikke evidens for at presumptivt feilaktig rapportering har noen effekt i denne bransjen.

Også når det gjelder "Godstrafikk på vei" gir Fisher-testen klar ikke-signifikans. For denne bransjen har vi kun data fra 2006.

Det er altså grunn til å tro at variabelen *presumtivt feilaktig rapportering* har en viss betydning for enkelte bransjer, men mindre betydning i andre.

5. Vurderingsvariablene på trinn 1 er viktige for sannsynligheten for avdekking

En viktig hensikt med vurderingsvariablene på trinn 1 er å tjene som en slags screening variable som kan tjene til å øke antall tilfeller av avdekking i det relativt lille utvalget som er realistisk å oppnå på trinn 2. Det er derfor viktig for senere datainnsamlinger at disse variablene faktisk betyr noe for avdekking (endrings) - sannsynligheten. Dette viser seg (heldigvis) å være tilfelle i det foreliggende materialet.

I utskrift 4, 5 og 6 ser vi på betydningen av vurderingsvariablene *MaxAvPoeng* (MAP) og *Samlet vurdering 811* (SV) i tillegg til B_1, B_2 og distriktsvariabelen *Sør*. Risikovariabelen *presumptivt feilaktig rapportering* (*I*) viser seg ikke å ha noen effekt i denne sammenhengen og er derfor fjernet. Utskrift 4 viser at MAP ikke synes å ha prediksjonskraft utover SV. Utskrift 5 og 6 viser at MAP og SV hver for seg har

[8] Har her benyttet samme definisjon av bedrifter med presumptivt korrekt rapportering og bedrifter med presumptivt feilaktig rapportering i pilot-dataene som er brukt i de foreliggende 2006-dataene. Opprinnelig var definisjonen litt annerledes for pilot-dataene. Dessuten var begrepsbetegnelsene noe annerledes. Se notat av P.G. Larsen , 10/1-2007.

signifikant prediksjonskraft. Fjerner man MAP fra utskrift 4 for å få utskrift 5, synker likelihooden ubetydelig, men fjerner man SV fra 4 for å få utskrift 6, synker likelihooden mer. Alt i alt synes derfor SV å inneholde noe mer informasjon (kanskje om revisors helhetsinntrykk som ikke er uttrykt i MAP) om endringssannsynligheten enn MAP. Det foreliggende trinn-2-utvalget er utelukkende basert på MAP, og det kan således reises spørsmål om SV kanskje ville vært bedre. På den annen side er forskjellen i prediksjonskraft ikke stor mellom de to variablene.

6. Andre kovariater's betydning for *I* og sannsynligheten for avdekking

I noen tilfeller vil en tilsynelatende ikke-signifikant forklaringsvariabel (som *I*) kunne få signifikant forklaringskraft ved å trekke inn andre kovariater. Dette ble (som i pilotstudien) prøvet for andre potensielle kovariater nevnt i avsnitt 1.

Det viste seg imidlertid at ingen av disse forklaringsvariable førte til en regresjonskoeffisient for *I* som er signifikant (eller nær signifikant) forskjellig fra null, og vi kan konkludere med at *I* kan elimineres som prediktor for avdekkings-sannsynligheten på trinn 2 i 2006.

En nærmere analyse, oppsummert nedenfor, tyder på at disse kovariatene, med unntak av *vurdering* diskutert i avsnitt 5, synes å ha lite vesentlig å bidra med når det gjelder forklaringskraft og prediksjonskraft utover B_2 , *Z* og *Sør* for avdekkings-sannsynligheten.

Vurderingsvariablene er her representert ved *Z* alene, både siden *Z* både bestemmer utvalget på trinn 2 og er sterkt korrelert med *samlet vurdering* (811). De øvrige potensielle kovariatene deles i to grupper, *fulle kovariater*, det vil si kovariater som har verdier for alle 83 enheter, og *begrensete kovariater* som inneholder manglende observasjoner ("missings"). De aktuelle fulle kovariatene utgjøres av følgende 19 variable:

Fulle kovariater:

(6.1) I, B1, B2, Z, EksternRegn, Ansatte, Øst, Sør, Vest, Midt, ENK, AS, Alder, komsentral, komnæring3, komnæring4, komnæring5, komnæring6, komnæring7

Dummien "komnæring2" er fjernet på grunn av for få observasjoner (2). Det betyr at næringsstruktur 1 (primærnæringskommuner) og 2 (blandede landbruks- og industrikommuner), er slått sammen til en kategori.

Begrensete kovariater (med manglende observasjoner) er angitt i tabell 6.1 med antall observasjoner:

Tabell 6.1 Potensielle begrensete kovariater

	Antall observasjoner	
	2005	2006
Bruttofortjeneste	43	24
Sum driftsinntekter	**73**	33
Leie av lokale	**65**	31
Driftsresultat	**73**	31
Næringsinntekt	**70**	30
Avgiftspliktig omsetning	**80**	**80**

For å unngå variable med for få observasjoner i regresjonskjøringene, velges følgende variable (uthevet i tabellen) som aktuelle begrensete kovariater:

Begrensete kovariater:

(2.2) Driftsinntekt05, Leielokal05, Driftsresultat05, Næringsinntekt05, Omsetning05, Omsetning06.

For ikke å tape for mye informasjon holdes i første omgang de begrensete kovariatene utenfor. I situasjoner med mange potensielle kovariater der det skal velges en blant et stort antall mulige forklaringsmodeller (og der det nesten alltid fins mange forskjellige submodeller som gir omtrent like god tilpasning til data i et begrenset datamateriale), er det vanlig først å sette opp en "full" bakgrunnsmodell som ramme med så mange potensielle forklaringsvariabler med i modellen som mulig, selv om mange av disse variablene kan være uten betydning. Denne fulle modellen, som i utgangspunktet antas sann, danner deretter bakgrunn og målestokk som aktuelle submodeller veies opp mot.

En rekke kriterier til å veie alternative submodeller kan benyttes. Vanlig er for eksempel å se på p-verdier for estimerte koeffisienter, likelihood-ratio (LR) testing, og diverse informasjonskriterier. De siste har den fordel at de også tar hensyn til kompleksiteten av submodellene (og "straffer" en modell med mange parametre). Og sist, men ikke minst, blant flere aktuelle kandidat-modeller som data ikke kan skille mellom, foretrekker man gjerne en som gir substansielt god mening. Av informasjonskriterier har jeg benyttet her de vanlige AIC, Akaike's informasjonskriterium, og hans Bayesianske modifikasjon, BIC.

Det er også viktig å være klar over at man bør ta de signifikante effektene som "oppdages" ved denne lete-strategien med en viss klype salt. Studier av slike lete-strategier har bl.a. vist at hvis man undersøker alle mulige submodeller (noe som kan være en formidabel oppgave i seg selv) så er det stor sjanse for at variable som faktisk har substansiell betydning, er med blant de signifikante variablene man oppdager, som er den gode nyheten. Den dårlige nyheten er man også med stor sannsynlighet vil oppdage spuriøst signifikante variable, dvs. variable uten substansiell betydning men som av rene tilfeldigheter blir signifikante i et begrenset datamateriale. Det er vanligvis mange av disse på grunn av det store antallet av mulige submodeller.

Tabell 6.2 viser resultatet av en slik prosess. To alternativer for full-modell er vist, den første med alle full-variablene, den andre med de 5 kommunenærings-variablene fjernet. Det viste seg nemlig at disse variablene praktisk talt aldri slo ut med signifikans. I tillegg er LR testobservator (lik differensen mellom deviansene) for testing av den reduserte fulle modellen mot den fulle, lik 5.84. Med 5 frihetsgrader gir dette en p-verdi på 0.32. Det er altså ingen evidens i data for at kommunenes næringsstruktur har betydning.

Tabell 6.2 Analyse basert på fulle kovariater.[9]

Avhengig Y	Full modell		Full modell minus Kommunenærings-var		Prediksjonsmodell		Alternativ prediksjonsmodell	
	Koeff.	P-verdi	Koeff.	P-verdi	Koeff.	P-verdi	Koeff.	P-verdi
I	-0.04989	0.950	-0.03821	0.958	-----	-----	-----	-----
B1	0.81168	0.511	0.05870	0.949	-----	-----	-----	-----
B2	-1.75535	0.117	-2.33656	0.024	-2.34956	0.006	-2.47607	0.004
Z	2.76477	0.007	2.16144	0.008	1.64169	0.009	1.90048	0.005
Ekst.Regn (R)	-0.77882	0.476	-0.90393	0.296	-----	-----	-----	-----
Ansatte	0.17684	0.134	0.02829	0.473	-----	-----	-----	-----
Øst	-2.06441	0.578	-0.08217	0.960	-----	-----	-----	-----
Sør	0.49756	0.893	2.44514	0.121	1.55626	0.016	2.06592	0.007
Vest	-1.51754	0.681	0.70479	0.690	-----	-----	-----	-----
Midt	-1.37086	0.679	1.09301	0.488	-----	-----	1.35604	0.134
ENK	1.51802	0.491	-0.15097	0.920	-----	-----	-----	-----
AS	0.29986	0.890	-0.44052	0.769	-----	-----	-----	-----
Alder	-0.04835	0.420	-0.02487	0.656	-----	-----	-----	-----
komsentral	-0.28875	0.491	-0.16821	0.531	-----	-----	-----	-----
komnæring3	-0.60479	0.765	-----	-----	-----	-----	-----	-----
komnæring4	-25.01450	0.991	-----	-----	-----	-----	-----	-----
komnæring5	-1.68421	0.489	-----	-----	-----	-----	-----	-----
komnæring6	-1.73741	0.556	-----	-----	-----	-----	-----	-----
komnæring7	-0.34298	0.888	-----	-----	-----	-----	-----	-----
Konstantledd	1.22091	0.725	-0.509062	0.804	-1.657749	0.002	-2.212449	0.001
Antall obs	83		83		83		83	
Log-likelihood	-28.2501		-31.1680		-34.1192		-32.9749	
Devians	56.5003		62.3360		68.2383		65.9497	
AIC	1.163		1.112		0.919		0.915	
BIC	-221.9		-238.1		-280.9		-278.7	

Det er naturlig å spørre seg om hvorfor man ikke simpelthen rapporterer den fulle modellen som resultatet av undersøkelsen og angir på bakgrunn av denne hvilke variable som synes å ha eller ikke ha betydning. Svaret er at vi vanligvis vil forvente en god del støy i en slik modell på grunn av at flere (kanskje til og med mange) av

[9] Fra utskrift 7 - 10 i appendiks 2

variablene som er med modellen ikke har noen reell betydning - noe som kan bidra til å skape høye p-verdier for variable som faktisk har betydning. I første full-modell ser vi at bare *Z* opptrer signifikant (både Z og B_2 i den reduserte full-modellen), men på grunn av formodentlig mye støy, vil vi forvente at flere betydningsfulle variable kan dekke seg under noen av de høye p-verdiene. Om vi skal ha håp om å avdekke disse, kan vi neppe unngå å se på submodeller med færre kovariater.

På den annen side, det at Z og B_2 kommer til syne allerede i full-modellen(e) med mye antatt støy, gir øket tyngde til dem som kandidater til forklaringsvariable av betydning. I tillegg viser det seg at de framstår som signifikante i de fleste (ikke alle) submodeller de er med i.

Jeg har ikke undersøkt alle mulige submodeller. Den første full-modellen har $2^{19} = 524\,288$ mulige submodeller, mens den reduserte har $2^{14} = 16\,384$. I stedet har jeg forsøkt en rekke trinnvise prosedyrer, forlengs, baklengs og blandet med varierende eksklusjons- og inklusjonskriterier samt sett på en rekke enkelttilfeller. De to siste modellene i tabellen er forslag til mulige kandidater til forenklet modell (som jeg har kalt prediksjonsmodeller). Blant alle kandidater jeg så på framsto disse to som "vinnerne" ut fra informasjonskriteriene AIC og BIC. AIC peker ut den fjerde modellen i tabell 6.2 med Z, B_2 , *Sør* og *Midt*, som "vinner", mens BIC peker ut den tredje, med bare Z, B_2 , *Sør* som kovariater. Ettersom AIC er kjent for en viss tendens til å ende opp med litt for store modeller, velger jeg den tredje modellen som mitt forslag til prediksjonsmodell,

(6.3) $$\text{logit}(p(x)) = \beta_0 + \beta_1 B_2 + \beta_2 Z + \beta_3 Sør$$

som estimert (fra utskrift 7 i appendiks 2 og tabell 6.2) blir

(6.4) $$\text{logit}(\hat{p}(x)) = -1,6577 - 2,3496 \cdot B_2 + 1,6417 \cdot Z + 1,5563 \cdot Sør$$

Sammenligner vi (6.3) med den fulle modellen ved en LR-test (likelihood-ratio-test), blir p-verdien 0,762, og sammenlignet med den redusert fulle blir p-verdien 0,880. Det er altså heller ikke ut fra denne synsvinkelen noe som tyder på at vi taper noe vesentlig ved å utelate de andre kovariatene.

Det er også av interesse å undersøke om de begrensete kovariatene gir grunn til å endre bildet. I tabell 6.3 har jeg estimert en modell der variabelen *presumptivt feilaktig rapportering* (*I*) samt de begrensete kovariatene er lagt til prediksjonsmodellen. Prediksjonsmodellen er re-estimert basert på det reduserte datasettet (61 observasjoner) for å kunne gjøre den sammenlignbar med modellen med de begrensete kovariatene. LR-testen av prediksjonsmodellen mot modellen med *I* og de begrensete kovariatene gir en p-verdi på 0,746. I tillegg ser vi at ingen av p-verdiene for koeffisientene til *I* og de begrensete kovariatene er nær signifikans. Det

er altså ingenting som tyder på at det er tilleggsinformasjon i *I* og de begrensete kovariatene for avdekkings-sannsynligheten utover prediktorene i modell (6.3).

Den siste modellen i tabell 6.3 er prediksjonsmodellen pluss *I* alene (estimert ut fra 83 observasjoner), som igjen bidrar til å underbygge konklusjonen om *I*'s manglende betydning.

Tabell 6.3 Resultater med begrensete kovariater[10]

Avhengig Y	Prediksjonsmodell pluss tilleggsvariable		Prediksjonsmodell (Redusert datasett)		Prediksjonsmodell pluss mangelfulle rapporter (I)	
	Koeff.	P-verdi	Koeff.	P-verdi	Koeff.	P-verdi
I	0.04213	0.963	-----	-----	-.2226125	0.726
B2	-3.04338	0.006	-2.28727	0.010	-2.352375	0.006
Z	1.54937	0.140	0.91911	0.220	1.689839	0.009
Sør	2.56865	0.016	1.82343	0.014	1.534327	0.018
Midt	1.59427	0.232	-----	-----	-----	-----
Driftsinntekt05	-3.13E-07	0.672	-----	-----	-----	-----
Leielokal05	-1.40E-06	0.845	-----	-----	-----	-----
Driftsresultat05	2.39E-06	0.660	-----	-----	-----	-----
Næringsinntekt05	-1.45E-06	0.802	-----	-----	-----	-----
Omsetning05	5.25E-07	0.426	-----	-----	-----	-----
Omsetning06	-1.11E-07	0.782	-----	-----	-----	-----
Konstantledd	-2.63469	0.030	-1.506444	0.018	-1.53883	0.014
Antall obs	61		61		83	
Log-likelihood	-21.7694		-24.3228		-34.0574	
Deviance	43.5388		48.6456		68.1149	
AIC	1.107		0.929		0.941	
BIC	-157.9		-185.7		-276.6	

Det kan også være av interesse å se på de estimerte sannsynlighetene basert på den endelige prediksjonsmodellen (6.3) som finnes i tabell 6.4.

[10] Fra utskrift 11 - 13 i appendiks 2

Tabell 6.4 Estimerte avdekkingssannsynligheter basert på prediksjonsmodell (6.3)

	Bransje	Funn trinn 1	Antall avdek-kinger trinn 2	Antall i alt	Relativ frekvens avdekking trinn 2	Estimert Avdekkings-sanns.h.	95% KI	
		Z=1	Y=1	n			Nedre	Øvre
Andre distrikter	51 Engroshandel	nei	1	9	0.111	0.160	0.063	0.349
		ja	1	2	0.500	0.496	0.272	0.722
	60 Godstransport	nei	1	19	0.053	0.018	0.003	0.103
		ja	0	2	0.000	0.086	0.015	0.371
	74 Renhold	nei	1	10	0.100	0.160	0.063	0.349
		ja	6	11	0.545	0.496	0.272	0.722
Sum			10	53				
Distrikt Sør	51 Engroshandel	nei	1	3	0.333	0.475	0.236	0.725
		ja	4	4	1.000	0.823	0.559	0.945
	60 Godstransport	nei	1	11	0.091	0.079	0.018	0.293
		ja	0	2	0.000	0.308	0.070	0.724
	74 Renhold	nei	4	7	0.571	0.475	0.236	0.725
		ja	2	3	0.667	0.823	0.559	0.945
Sum			12	30				
Total sum			22	83				

Vi oppsummerer:

- Det er større estimert avdekkingssannsynlighet i distrikt Sør (Vestfold, Buskerud, Telemark, Vest-Agder) enn de andre distriktene (oddsfaktor: 4,7).
- Det er større estimert avdekkingssannsynlighet blant virksomheter som gir ”funn” på trinn 1 (Z =1) (oddsfaktor: 5,2). Dette innbærer at ”screening”s-aktiviteten på trinn 1 for å øke avdekkingssannsynligheten på trinn 2 synes å ha fungert tilfredstillende.
- Det er evidens for at bransje 60 (godstransport) har generelt lavere avdekkingssannsynlighet enn de andre bransjene i de observerte distriktene (oddsfaktor: 0,1).
- Selv om den valgte prediksjonsmodellen har redusert usikkerheten en god del i forhold til de observerte relative frekvensene i tabell 6.4, så viser konfidensintervallene at det er betydelig usikkerhet igjen i de estimerte avdekkingssannsynlighetene - noe som ikke er uventet sett i lys av det tross alt lille utvalget som foreligger.
- Det er lite som tyder på at de øvrige kovariatene i studien har vesentlig informasjon for avdekkingssannsynlighetene utover de kovariatene som er tatt med i prediksjonsmodellen. Imidlertid, her er det ikke-signifikante sammenhenger det siktes til. Det er viktig å være klar over at det at en variabel ikke er signifikant i en regresjonssammenheng ikke nødvendigvis betyr at den ikke betyr noe - særlig på grunnlag av et begrenset datamateriale. Det kan godt være at den ikke betyr noe, men det kan også være at den faktisk betyr noe, men at begrenset informasjon i data

ikke har kunnet oppdage det. Det kan derfor være nyttig også å studere deskriptive og ikke-signifikante sammenhenger i data som har potensiell substansiell interesse med henblikk på en eventuell senere undersøkelse.

- Den initiale risikovariabelen, *presumptivt feilaktig rapportering* (*I*), synes ikke å ha betydelig informasjon utover funn-variabelen på trinn 1, *Z* når det gjelder å predikere avdekking på trinn 2. Noe informasjon foreligger imidlertid som tyder på at dette kan variere mellom bransjer. I tillegg kan det være enkeltelementer som inngår i *I* som kan være viktig. (Jfr. avsnitt 4 og 7).

7. Har de enkelte risiko faktorene som inngår i *I* betydning?

De fire enkelte risikofaktorene som er med eller har vært med i *I*, er

1. Manglende og uriktige oppgaver
2. Betalingsanmerkninger eller manglende proveny
3. Registeropplysninger
4. Omsetning innenfor flere satser

Tabell 7.1 Krysstabell for Manglende og uriktige oppgaver

			Endringer - alle typer (*Y*)		Total
			Nei (Y = 0)	Ja (Y = 1)	
Manglende og uriktige oppgaver	Nei	Count	57	16	73
		%	78,1%	21,9%	100,0%
	Ja	Count	4	6	10
		%	40,0%	60,0%	100,0%

Sannsynligheten for $Y = 1$ når risikofaktor 1 ikke er tilstede estimeres til 0,22, mens den tilsvarende sannsynligheten når 1 er tilstede estimeres til 0,60. I henhold til Fishers eksakte test (se avsnitt 4) er denne forskjellen signifikant (p-verdi 0,019). Den tilsvarende forskjellen i estimert sannsynlighet for de andre tre risikofaktorene er imidlertid langt fra signifikante med p-verdier godt over 0,5. Det foreligger altså evidens for at faktor 1 har betydning, mens de andre tre ikke har det. Det bør tas et forbehold når det gjelder konklusjonen for faktorene 2,3,4 på grunn av små datasett.

8. Teoretisk grunnlag for analyse av endringstallene

Utvalget er stratifisert over mange strata (66) med svært små utvalg i hvert stratum (fra 1 til 5). På grunn av bortfall er noen strata faktisk uten observasjoner (jfr. tabell 18 i appendiks 2). Dette betyr at standard design-basert analyse er uegnet for denne situasjonen. Min tilnærming vil være såkalt modell-basert i stedet som gir større muligheter for å utnytte homogenitetstrekk og imputering i utvalget.

I tillegg er vi nødt til å kontrollere for den intenderte skjevheten i trinn-2- resultatene foråsaket av screeningsvariabelen *Z*. Dette gjøres vanligvis ved å modellere betingete sannsynlighetsfordelinger for responsene (*X* og *Y*) der *Z* inngår blant variablene det betinges med hensyn på. Selve skjevhetskontrollen består da i å senere aggregere ut *Z*

fra de betingete fordelingene. Skjevhet kan også oppstå via "mangel-mekanismen", det vil si via alle de virksomhetene i trinn-1-utvalget som ikke har observervasjoner av *X* og *Y*. I appendiks **A1.2** blir det påvist at mangel-mekanismen i dette tilfellet har egenskapen "missing at random" (MAR), som garanterer at denne typen av skjevhet ikke gjør seg gjeldende.

La $k = (a,b,l)$ betegne stratum *k* der $a = 1,2,\ldots,11$ betegner fylke, b = 1,2,3 bransje og l = 1, 2 gruppene *presumptivt korrekt rapportering* (I = 0) og *presumptivt feilaktig rapportering* (I = 1) h.h.v.

For en gitt bransje har vi $K = 22$ strata som nummereres $k = 1,2,\ldots,22$. I det følgende undertrykkes indeksen *b* for bransje. La

$N_k =$ antall virksomheter i stratum *k*.
$n_k =$ utvalgsstørrelsen i stratum *k*.
$X_{ki} =$ størrelsen på endringsbeløpet ("total endring", X_2, i avsnitt 1) for virksomhet *i* i stratum *k*, $i = 1,2,\ldots,N_k$. (De fleste X_{ki} vil være 0.)

Vi ønsker bl.a. å anslå bransje-totalen

$$T = \sum_{k=1}^{K} \sum_{i=1}^{N_k} X_{ki} \tag{8.1}$$

Ved den klassiske design-baserte tilnærmingen antas X_{ki} å være faste tall mens det er indeksen *i* som blir stokastisk i utvalget og danner grunnlaget for estimering og inferens. Hvis $\tilde{X}_k$ betegner gjennomsnittlig *X* i stratum *k*, og $\bar{X}_k$ gjennomsnittet i utvalget fra stratum *k*, blir totalen estimert ved

$$T = \sum_{k=1}^{K} N_k \tilde{X}_k \approx \sum_{k=1}^{K} N_k \bar{X}_k = \hat{T}$$

der $\bar{X}_k$ estimerer $\tilde{X}_k$ (forventningsrett).

Ved den modell-baserte tilnærmingen antas tvert i mot at X_{ki} er stokastiske variable (også i populasjonen), mens indeksen *i* antas faste tall i utvalget. Det design-baserte uttrykket

" $X_{ki_1}, X_{ki_2}, \ldots, X_{ki_{n_k}}$ er et rent tilfeldig utvalg fra stratum *k*"

oversettes til uttrykket

" $X_{k1}, X_{k2}, \ldots, X_{kn_k}$ er uavhengige og identisk fordelte med fordeling $f_k(x)$ "

Dermed blir problemet å estimere det faste tallet $T = \sum_{k=1}^{K} \sum_{i=1}^{N_k} X_{ki}$ oversatt til problemet å predikere den stokastiske variabelen $T = \sum_{k=1}^{K} \sum_{i=1}^{N_k} X_{ki}$.

Den "beste" prediktoren (i betydningen den som minimerer forventet kvadratisk prediksjonsfeil) er som kjent forventningen, $E(T)$, eller et estimat for denne dersom den er ukjent. Vi har

$$E(T) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} E(X_{ki}) = \sum_{k=1}^{K} N_k E(X_k)$$

der X_k er en stokastisk variabel med fordeling $f_k(x)$ (som vi skal modellere).
Vår prediktor for *T* blir således

(8.2) $$\hat{T} = \sum_{k=1}^{K} N_k \hat{E}(X_k)$$

og oppgaven er redusert til å finne en fornuftig estimator for $E(X_k)$.
Vi trenger også et anslag på variansen

(8.3) $$\text{var}(T) = \sum_{k=1}^{K} N_k \text{var}(X_k)$$

Til hjelp for dette trekker vi inn de andre kovariatene. La U betegne vektoren av aktuelle kovariater som i utgangspunktet antas eksogene. Forventning og varians for X_k kan nå knyttes til *U* via de berømte setningene for dobbelt forventning:

(8.4) $$E_k(X) = E_k E_k(X \mid U)$$

(8.5) $$\text{var}_k(X) = E_k\left[\text{var}_k(X \mid U)\right] + \text{var}_k\left[E_k(X \mid U)\right]$$

der jeg har plassert indeksen *k* på "*E* " og "var" i stedet - for å understreke at forventning og varians beregnes ut fra fordelingen $f_k(x)$.

Siden utvalget på trinn 2 delvis er bestemt av en post-stratifisering basert på screening-variabelen *Z*, er det uhensiktsmessig å modellere direkte fordelingen for $(X \mid U)$ - dvs. den betingete fordelingen for *X* gitt *U* i stratum *k*. I stedet ser jeg på fordelingen for $(X \mid U, Z, Y)$, der *Y* er tatt med i betingelsen for å kunne kontrollere for 0-observasjonene for *X* (ettersom $Y = 0 \quad \Rightarrow \quad X = 0$).

De 22 positive endringstallene synes klart å indikere en høyreskjev fordeling. To naturlige kandidater for modellering av fordelingen for $(X \mid U, Z, Y = 1)$ er derfor log-normal fordeling eller gammafordeling med log-lineær linkfunksjon. De to fordelingene ligner på hverandre og er ofte vanskelig å diskriminere ut fra data. Jeg har valgt gammafordelingen her først og fremst siden den har visse tolkningsfordeler i og med at den ikke forutsetter noen log-transformasjon av responsen $X > 0$. En residualanalyse (jfr. fig. 9.1) viser seg heller ikke å gi sterke argumenter mot denne antakelsen. På den annen side er det klart at den log-normale fordelingen representerer et reelt alternativ, og det kunne vært en ide å gjennomføre analysen med begge alternativer - noe jeg ikke har gjort.

I likhet med logistisk regresjon er også modellen med gammafordeling et spesialtilfelle av GLM (generalisert lineær modellering) der det foreligger en utstrakt litteratur og der det er utviklet effektive beregningsalgoritmer. I GLM-litteraturen parametriseres gammafordelingen gjerne som følger. En variabel X er gammafordelt med parametre (μ, φ) som er definert slik at

$$E(X) = \mu \quad \text{og} \quad \text{var}(X) = \varphi\mu^2$$

Modellen impliserer en konstant variasjonskoeffisient for X, $\sqrt{\text{var}(X)}\,/E(X) = \sqrt{\varphi}$, som innebærer at standardavviket for X er proporsjonalt med forventningen. Avhengigheten til kovariatene legges inn i forventningen μ. For å unngå eventuelt negative estimater for forventet endringstall, velger jeg en log-lineær link-funksjon, $\eta = \ln(\mu)$, der η spesifiseres som en lineær funksjon av kovariatene. Parameteren φ kalles ofte i GLM-litteraturen for en dispersjonsparameter. STATA kaller den en scale-parameter.

(8.6) **Den foreslåtte modellen for** $(X \mid U, Z, Y)$ **er**

(**i**) $(X \mid U, Z, Y = 0) = 0$ med sannsynlighet 1

(**ii**) $(X \mid U, Z, Y = 1)$ er gammafordelt med parametre (μ_z, φ), der

$$\mu_z = E(X \mid U, Z = z, Y = 1) = e^{\eta} = e^{\gamma' U + \delta z}$$

$$\text{var}(X \mid U, Z = z, Y = 1) = \varphi\mu_z^2$$

(8.7) **Modeller for** ***Y*** **og** ***Z***. Her spesifiseres logistiske regresjonsmodeller, som ved hjelp av den inverse logit-funksjonen, $\text{logit}^{-1}(x) = \dfrac{e^x}{1 + e^x}$, kan uttrykkes

(**i**) $p_z = P(Y = 1 \mid U, Z = z) = \text{logit}^{-1}(\beta' U + \xi z)$

(**ii**) $q = P(Z = 1 | U) = \text{logit}^{-1}(\alpha' U)$

I alle regresjonsuttrykkene omfatter parametervektorene α, β, γ konstantledd. Merk også at μ_z og p_z er funksjoner av *U* og *Z*, mens *q* er en funksjon av *U* alene.

Spesifikasjonene (8.6) og (8.7) bestemmer den simultane fordelingen for responsen (X, Y, Z) gitt *U*.

Følgende resultater er utledet i appendiks A1.1:

(8.8) **Resultater for responsen** (X, Y, Z)

(**i**) Sett $\overline{p} = P(Y = 1 | U)$. Da gjelder

$$\overline{p} = P(Y = 1 | U) = p_0(1 - q) + p_1 q \text{, der}$$

$$p_1 = P(Y = 1 | U, Z = 1) = \text{logit}^{-1}(\gamma' U + \xi)$$
$$p_0 = P(Y = 1 | U, Z = 0) = \text{logit}^{-1}(\gamma' U)$$

Anta at $\mu_z = E(X | U, Y = 1, Z = z)$ ikke avhenger av *z*. Sett $\mu = \mu_0 = \mu_1$. Da gjelder[11]

(**ii**) $E_k(X | U) = \mu \overline{p}$

(**iii**) $\text{var}_k(X | U) = \varphi \mu^2 \overline{p} + \mu^2 \overline{p}(1 - \overline{p})$

(**iv**) $E_k X = E_k(\mu \overline{p})$

(**v**) $\text{var}_k X = (1 + \varphi) E_k\left(\mu^2 \overline{p}\right) - \left[E_k \mu \overline{p}\right]^2$

Merk at avdekkingssannsynligheten $E_k \overline{p} = E_k\left(P(Y = 1 | U)\right)$, dvs gjennomsnittlig $P(Y = 1 | U)$ over de variable i *U* som varierer i stratum *k*, kan tolkes som et mål på prevalens (utbredelse av endringstilfeller) i stratum *k* siden $P(Y = 1 | U)$ representerer avdekkingssannsynligheten kontrollert for screening-variabelen *Z*. Estimater for denne kan finnes i tabell 10.2.

I appendiks A1.2 er den simultane likelihood-funksjonen satt opp og diskutert. Det blir vist at den kan skrives som et produkt av tre faktorer med forskjellige

[11] Mer generelle uttrykk uten denne forutsetningen er gitt i appendiks A1.1.

parametersett som inngår i μ, p_z og q henholdsvis. Dette innebærer at vi uten tap av informasjon kan estimere og analysere de tre elementene separat samt at maximum likelihood estimatorene (ML) fra de tre elementene er (asymptotisk) uavhengige. Dette forenkler analysen betraktelig i og med at datagrunnlaget for de tre elementene er forskjellig. For μ består datagrunnlaget av 22 observasjoner, for p_z består det av 83 observasjoner og for *q* av 277 observasjoner. Vi kan også slutte at analysen for p_z gjennomført i avsnitt 5 og 6 kan anses som en full ML-analyse. Analysen av μ og *q* er gjennomført i avsnitt 9.

9. Betydning av andre kovariater for sannsynligheten for funn på trinn 1 og for forventet endringstall ($X = X_2$) gitt avdekking

Her har det blant annet interesse studere i hvilken grad kovariater *ikke* har betydning for endringstallene ettersom dette vil lette aggregering over strata.

9.1 Forventet endringstall gitt avdekking. I tabell 9.1 har jeg gjennomført en tilsvarende analyse for de positive endringstallene (*X*) som for avdekkingssannsynligheten i avsnitt 6, basert på modellen i (8.6) **ii.** Motivasjonen er den samme som i avsnitt 6 og forbeholdene likeså. Jeg starter med en full modell der jeg har tatt med alle variable fra avsnitt 1 som ikke inneholder manglende observasjoner. De seks dummiene for næringsstruktur av kommunen har jeg erstattet med en dummy, *Industri*, som indikerer en industrikommune (kategori 3 eller 5). Dette fordi 6 dummy variable for noe som i realiteten er én variabel blir for kostbart på bakgrunn av bare 22 observasjoner.

Tabell 9.1 GLM-analyse av forventet endringstall gitt avdekking
(Fra utskrift 14 og 15 i app.2)

Avhengig X	Full modell		Prediksjonsmodell	
	Koeff.	P-verdi	Koeff.	P-verdi
Z	0.463610	0.827	----	----
B1	-0.513197	0.715	----	----
B2	0.919541	0.538	----	----
Midt	0.327008	0.910	----	----
Øst	-2.462447	0.124	-1.753912	0.006
Sør	-0.889017	0.667	----	----
Vest	-1.477969	0.499	----	----
AS	0.515235	0.608	----	----
R (ekst. regnsk.fører)	-1.961653	0.040	-1.779677	0.001
I	0.181894	0.880	----	----
Komsentral	0.365308	0.333	0.333767	0.005
Omsetn06	0.000000	0.800	----	----
Alder	-0.065096	0.462	----	----
Industri	0.929301	0.433	----	----
Konstantledd	11.733670	0.000	11.657900	0.000
Antall obs	22		22	
Dispersjonsparam. (φ)	1.3248		0.6916	
Log-likelihood	-285.3740		-288.1268	
Devians	18.2641		23.7698	
AIC	27.3067		26.5570	
BIC	-3.3732		-31.8690	

Ved en tilsvarende søkestrategi som i avsnitt 6 endte jeg opp med følgende forslag til redusert modell (prediksjonsmodell) angitt i de to siste kolonnene i tabellen (som impliserer at $\mu_z = \mu$ ikke avhenger av *z*) :

(9.1) $$\ln\left[E(X \mid U, Z, Y = 1)\right] = \ln(\mu) = \gamma_0 + \gamma_1 \text{Øst} + \gamma_2 \text{Komsentral} + \gamma_3 \text{R}$$

som estimert blir

(9.2) $$\ln\left[\hat{E}(X \mid U, Z, Y = 1)\right] = 11{,}658 - 1{,}754 \cdot \text{Øst} + 0{,}334 \cdot \text{Komsentral} - 1{,}780 \cdot \text{R}$$

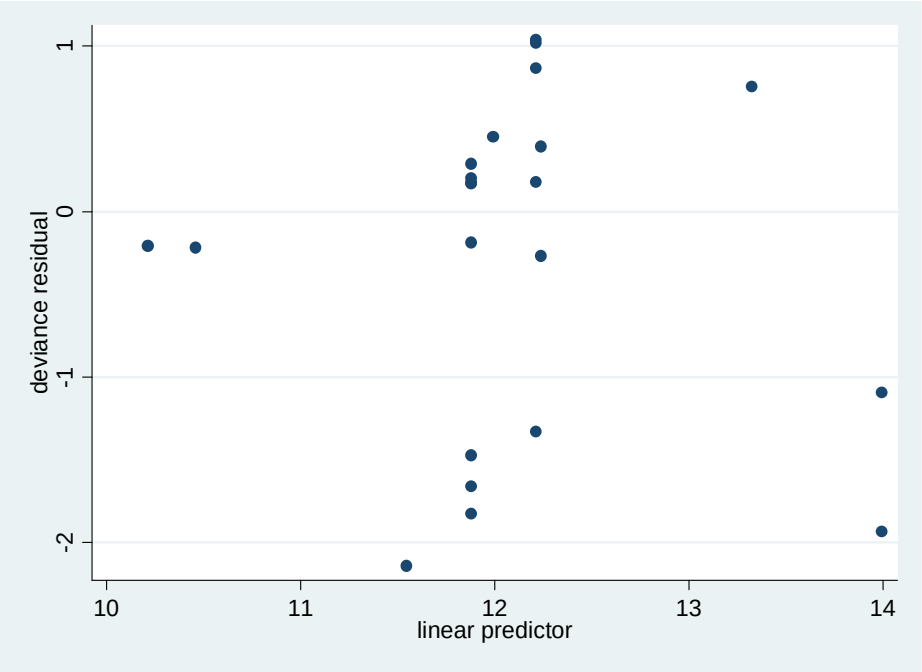
der estimatene er gitt i tabell 9.1. En likelihood-ratio (LR) test av den reduserte modellen mot den fulle gir en p-verdi på 0,904, som indikerer at de øvrige kovariatene ikke synes å bidra til forklaring utover de tre variablene som er tatt med. Her må det tas samme forbehold som i avsnitt 6 siden den reduserte modellen ikke er valgt ut a priori men på grunnlag av å lete i data.

Noen iøynefallende konklusjoner byr seg:

- Screeningvariabelen, funn på trinn 1 (*Z*), synes ikke å ha betydning for endringstallenes størrelse gitt avdekking. Dette styrkes ved at *Z* ikke var signifikant i noen av de mange sub-modellene for endringstallet jeg så på der *Z* var med (jfr. tabell 9.1). Sammen med resultatene i avsnitt 6 tyder dette på at *Z* har betydning for endringstallet, men bare via sannsynligheten for at endring foreligger (funn på trinn 1 øker sannsynligheten), og ikke på hvor stor denne endringen er.
- Det omvendte forhold synes å gjøre seg gjeldende for variabelen "Ekstern regnskapsfører" (*R*). Etter avsnitt 6 synes ikke *R* å ha noe å si for sannsynligheten for at endring foreligger (i.e., *R* er ikke signifikant i tabell 6.2). På den annen side, etter tabell 9.1, synes *R* å være viktig for størrelsen på endrings-beløpet gitt avdekking. *R* var den eneste variabelen som kom ut signifikant i den fulle modellen for endringsbeløpets størrelse, som er et sterkt resultat i seg selv, og som indikerer at den bør være med i enhver submodell som foreslås. En negativ koeffisient tilsier at tilstedeværelse av ekstern regnskapsfører fører til en tendens til redusert gjennomsnittlig endringsbeløp der endring er aktuelt.
- Det er også interessant at bransje ikke synes å ha noen effekt på endringsbeløpets størrelse gitt avdekking. Fra avsnitt 6 synes bransje 2 (godstrafikk på vei) å ha en mindre sannsynlighet enn de andre bransjene for å påavdekkinge et tilfelle av avdekking, $X > 0$, men gitt avdekking, er det ikke noe som tyder på at endringsbeløpet i snitt er forskjellig fra de andre bransjene.
- Det kan virke ut fra prediksjonsmodellen at endringsbeløpet gitt avdekking i snitt er lavere i distrikt Øst enn ellers. Den negative koeffisienten for Øst kompenseres noe av at de fleste kommunene i Øst er sentrale med høy verdi på variabelen *Komsentral* som har en positiv regresjonskoeffisient. Likevel, til tross for dette, indikerer resultatene i tabell 10.2 et lavere gjennomsnittlig endringsbeløp i øst.

I figur 9.1 har jeg gitt et residualplott for den foreslåtte prediksjonsmodellen for μ, der devians-residualene er plottet mot den lineære prediktoren $\hat{\eta} = \ln(\hat{\mu})$. I GLM-litteraturen er en rekke forskjellige residualtyper diskutert, men det er som oftest devians-residualene som anbefales. Plottet gir ingen sterke tendenser som taler i mot prediksjonsmodellen basert på gamma-fordelingen, bortsett kanskje fra en svak tendens til venstreskjevhet i residual-fordelingen. Dette kan godt skyldes den ene ekstreme *X*-observasjonen til en rengjøringsbedrift fra Nord på 1,2 millioner (de andre varierer i området 0 - 500 000), som kanskje har trukket opp estimatet for η mer enn det som ville skjedd i et større materiale dersom denne *X*-observasjonen er mer ekstrem enn det som vil være vanlig i den aktuelle gamma-fordelingen. På den annen siden er materialet altfor lite til å kunne ta stilling til dette spørsmålet i og med at vi bare har en ekstrem observasjon - som på den annen side ikke er så ekstrem at den ville utelukke gamma-fordelingen.

Figur 9.1 Residual plott for prediksjonsmodellen for *X* gitt avdekking i (9.1)



9.2 Prediksjonsmodell for funn på trinn 1 (Z). Resultatet av en tilsvarende analyse som for μ og p_z er gitt i tabell 9.2. Samme forbehold som beskrevet i avsnitt 6 gjelder også her selv om data-grunnlaget her er større (277 observasjoner). I den fulle modellen er alle dummiene for kommunens næringsstruktur tatt med, men synes ikke å ha betydning for prediksjon av *Z*. En LR-test av den foreslåtte prediksjonsmodellen mot den fulle modellen gir en p-verdi på 0,294, slik at det er ingen ting som tyder på at de øvrige kovariatene har betydning for prediksjon av *Z* utover de som inngår i prediksjonsmodellen (med et lite forbehold knyttet til at prediksjonsmodellen er valgt ut fra data og ikke a priori).

Tabell 9.2 Logistisk regresjons-resultater for funn på trinn 1 (screening)
(Fra utskrift 16 og 17 i app.2)

Avhengig Z	Full modell		Prediksjonsmodell	
	Koeff.	P-verdi	Koeff.	P-verdi
I	-0.5816859	0.201	----	----
B1	-1.0928930	0.081	-1.199268	0.017
B2	-2.0169950	0.001	-2.083633	0.000
R (ekst.regnsk.fører)	0.4874139	0.436	----	----
Ansatte	-0.2163971	0.077	-0.279384	0.008
Øst	0.1282347	0.855	----	----
Sør	-2.3985830	0.006	-1.537591	0.009
Vest	-1.7176540	0.047	-1.266519	0.059
Midt	-1.3458700	0.087	----	----
ENK	0.4575708	0.659	----	----
AS	-0.2921491	0.781	----	----
Alder	-0.0529451	0.085	----	----
Komsentral	0.1139704	0.567	----	----
Komnaer2	-2.5843280	0.167	----	----
Komnaer3	-1.3576660	0.376	----	----
Komnaer4	-16.8089500	0.987	----	----
Komnaer5	-1.8585770	0.263	----	----
Komnaer6	-1.6823340	0.293	----	----
Komnaer7	-1.6702220	0.328	----	----
Konstantledd	1.4825400	0.356	-0.203599	0.555
Antall obs	276		277	
Log-likelihood	-72.5090		-80.6698	
Devians	145.0179		161.3397	
AIC	0.6704		0.6258	
BIC	-1293.8050		-1362.7690	

Den foreslåtte prediksjonsmodellen for Z blir etter dette

(9.3) $$\text{logit}(q) = \text{logit}(P(Z=1|U)) = \alpha_0 + \alpha_1 B_1 + \alpha_2 B_2 + \alpha_3 \text{Sør} + \alpha_4 \text{Vest} + \alpha_5 \text{Ansatte}$$

som estimert blir

(9.4) $$\text{logit}(\hat{q}) = -0,204 - 1,199 \cdot B_1 - 2,084 \cdot B_2 - 1,538 \cdot \text{Sør} - 1,267 \cdot \text{Vest} - 0,279 \cdot \text{Ansatte}$$

10. Estimering av total sum endringstall pr. bransje

Etter analysen i avsnitt 6 og 9 blir den endelige vektoren av andre kovariater, U, (som antas eksogen) redusert til

(10.1) $U = (B_1, B_2, \text{Øst, Sør, Vest, } R(\text{Ekstern regnskapsfører}), \text{ Ansatte, Komsentral})$

Det er disse 8 variablene som synes å ha innvirkning, direkte eller indirekte, på fordelingen for $(X, Y, Z \mid U)$ bestemt av (8.6) - (8.8), mens de øvrige kovariatene i avsnitt 1 ikke synes å ha vesentlig betydning. Som nevnt ovenfor, et forbehold som likevel alltid gjelder når kovariater velges ut fra mange kandidater ved hjelp av data, er muligheten for andre kombinasjoner av forklaringsvariable som kan gi omtrent like god modelltilpasning.

For en gitt bransje er oppgaven nå å predikere totalen (bransje-indeksen *b* undertrykkes):

(10.2) $$T = \sum_{k=1}^{K} \sum_{i=1}^{N_k} X_{ki}$$

med prediktoren (jfr. avsnitt 8)

(10.3) $$\hat{T} = \hat{E}(T) = \sum_{k=1}^{K} N_k \hat{E}_k(X)$$

I følge (8.8) er $E_k X = E_k(\mu\bar{p})$. Her er $\mu\bar{p}$ en funksjon $\mu\bar{p} = h(U;\theta)$ der $\theta' = (\alpha', \beta', \gamma', \varphi)$ er parametervektoren. La indeksen *i* betegne enhet *i* i stratum *k*. Skriv så $h(U_i;\theta) = \mu_i \bar{p}_i$. På grunn av den stratifiserte utvalgsdesignen som er brukt får vi (eksakt ut fra en design-basert synsvinkel):

(10.4) $$E_k(\mu\bar{p}) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mu_i \bar{p}_i = \frac{1}{N_k} \sum_{i=1}^{N_k} h(U_i;\theta)$$

som estimeres ved

(10.5) $$\hat{E}_k(\mu\bar{p}) = \frac{1}{N_k} \sum_{i=1}^{N_k} h(U_i;\hat{\theta})$$

Når jeg tillater meg å bruke en design-basert synsvinkel her skyldes det at vi ikke er interessert i å modellere variasjonen av *U* (dvs. for de tre variablene, *R*, Ansatte, Komsentral) utover den perioden vi studerer, og for denne perioden er variasjonen til *U* helt kjent via skattedirektoratets database for de aktuelle strataene. Dermed er usikkerheten ved estimering av $E_k(\mu\bar{p})$ (og dermed $E(T)$) redusert til estimeringsusikkerheten for $\hat{\theta}$ alene. Formler for (asymptotisk) standardfeil for $\hat{E}_k(\mu\bar{p})$ er i henhold til dette utledet i appendiks A1.3.

Dersom vi altså utnytter skattedirektoratets database for de aktuelle strataene, blir prediktoren

(10.6) $$\hat{T} = \hat{E}(T) = \sum_{k=1}^{K} N_k \hat{E}_k(X) = \sum_{k=1}^{K} \sum_{i=1}^{N_k} h(U_i;\hat{\theta}) \quad \text{der} \quad h(U_i;\hat{\theta}) = \hat{\mu}_i \hat{\bar{p}}_i .$$

Dersom vi (som i denne rapporten) som tilnærmelse bare utnytter informasjonen i det opprinnelige utvalget på trinn 1 med n_k observasjoner i stratum *k*, blir prediktoren

(10.7) $$\hat{T} = \hat{E}(T) = \sum_{k=1}^{K} N_k \hat{E}_k(X) = \sum_{k=1}^{K} \frac{N_k}{n_k} \sum_{i=1}^{n_k} h(U_i;\hat{\theta})$$

Merk også at valget av kovariatvektor *U* i (10.1) representerer en betydelig aggregering av de 22 strataene for en gitt bransje ned til 4 strata gitt ved distriktene *Øst, Sør, Vest* og *Midt/Nord* slått sammen. Det betyr at *K* i (10.6) og (10.7) er lik 4, og der N_k, n_k betegner tilsvarende aggregerte stratum- og utvalgsstørrelser.

Resultatet av bruk av (10.7) er gitt i tabell 10.1.

Tabell 10.1 Aggregerte endringstall (*T*) predikert pr. bransje og distrikt (enhet million kroner).

Bransje	Distrikt	Stratum størrelse	Predikert total T	Estimert st.avvik T	Standardfeil på estimert ET	Total prediksjons-standardfeil
51	Øst	416	8.19	1.05	4.83	4.95
Engroshandel	Sør	295	73.99	6.72	36.88	37.48
klær, sports-	Vest	229	13.88	3.10	7.38	8.01
og fritidsutstyr	Midt/Nord	116	8.16	2.52	4.18	4.88
51 Total		1056	104.22	7.88	46.34	47.01
60	Øst	2066	2.75	0.58	14.61	14.62
Godstransport	Sør	1522	40.41	5.08	98.41	98.54
på vei	Vest	1170	10.89	3.23	21.26	21.50
	Midt/Nord	1140	3.66	1.08	29.37	29.39
60 Total		5898	57.71	6.14	162.12	162.24
74	Øst	438	9.48	1.12	5.22	5.34
Rengjøring	Sør	308	44.77	4.61	17.37	17.97
	Vest	232	16.98	3.61	8.93	9.63
	Midt/Nord	188	10.61	2.25	4.08	4.66
74 Total		1166	81.84	6.37	28.33	29.04

Prediksjons-usikkeheten er angitt ved prediksjons-standardfeilen (PSE) i siste kolonne. Den består av to komponenter, standardavviket for den stokastiske totalen *T*, SD(*T*), estimert i femte kolonne på grunnlag av (8.3) og (8.8)**v**, og standardfeilen ved estimering av *E*(*T*), $\text{SE}(\hat{E}(T))$, basert på formler utviklet i appendiks A1.3. Sammenhengen er

$$(10.8) \qquad \mathrm{PSE}(\hat{T}) = \sqrt{E\left[\left(T - \hat{T}\right)^2\right]} = \sqrt{\left(\mathrm{SD}(T)\right)^2 + \left(\mathrm{SE}(\hat{E}(T)\right)^2}$$

Vi ser av tabell 10.1 at usikkerheten ved totalanslagene her er betydelig og sterkt dominert av estimeringsusikkerheten i $\hat{E}(T)$. Dette bør ikke være overraskende sett i lys av det tynne datagrunnlaget for ($X > 0$)-responser (22 observasjoner med betydelig variasjon seg imellom, jfr. kolonne 7 i tabell 10.2).

Alle anslag i tabell 10.1 er basert på prediktoren i (10.7) og gjennomsnitt basert på det opprinnelige utvalget på trinn 1. En viss forbedring av kvaliteten av anslagene kan oppnås ved i stedet å benytte prediktor (10.6) og fullstendige gjennomsnitt basert på skattedirektoratets database. Dette vil imidlertid ikke ha vesentlig betydning for nivået på usikkerhets-anslagene ettersom jeg her har ignorert den delen av usikkerheten som skyldes å ha erstattet diverse stratumgjennomsnitt med tilsvarende gjennomsnitt fra trinn-1-uvalget.

Tabell 10.2 viser hvordan usikkerheten kommer til uttrykk i de enkelte (aggregerte) strataene ved sammenligning av avdekking-frekvenser og gjennomsnittlig endringsbeløp gitt avdekking, med tilsvarende estimerte størrelser.

Tabell 10.2 Observert og estimert avdekking-frekvens og gjennomsnittlig endringsbeløp i de aggregerte strataene

Bransje	Distrikt	Utvalg trinn 1	Utvalg trinn 2	Antall avdekking	Observert $\overline{X}$ gitt avdekking	Observert st.avvik gitt avdekking	Estimert $E(\overline{X} \mid \text{treff})$	Estimert sannsynlighet for avdekking
51	Øst	18	3	0	----	----	96 682	0.205
Engroshandel	Sør	28	7	5	176 009	185 034	515 709	0.487
Klær, sports-	Vest	12	2	0	----	----	351 976	0.175
og fritidsutstyr	Midt/Nord	16	6	2	124 682	170 670	344 504	0.216
51 Total		74	18	7	161 344	168 247	350 215	0.309
60	Øst	21	8	0	----	----	57 608	0.023
Godstransport	Sør	46	13	1	190 545	---	317 800	0.083
på vei	Vest	16	2	0	----	----	483 259	0.019
	Midt/Nord	37	11	1	488 785	----	143 089	0.022
60 Total		120	34	2	339 665	210 888	240 458	0.045
74	Øst	15	7	3	161 148	135 820	79 477	0.272
Rengjøring	Sør	35	10	6	145 635	121 633	287 154	0.510
	Vest	18	6	2	251 000	253 144	385 992	0.204
	Midt/Nord	29	8	2	611 000	832 972	211 102	0.256
74 Total		97	31	13	237 020	318 366	250 643	0.340
Grand Total		291	83	22				0.211

Det er tilsynelatende store avvik mellom observert $\bar{X}$ gitt avdekking og estimert $E(\bar{X} \mid \text{treff})$, det mest ekstreme for distrikt Sør i bransje 51. På den annen side er utvalgene bak $\bar{X}$ (se kolonne 5) svært små. For å få et perspektiv på om avvikene virker urimelig store, kan vi gjennomføre en rent hypotetisk illustrerende beregning i analogi med de fem observasjonene i Sør for bransje 51:

Anta $X_1, X_2, \ldots, X_5$ er uavhengige og normalfordelte med forventning μ og kjent standardavvik, $\sigma = 185000$. Anta videre at vi ønsker å predikere en ny observasjon av $\bar{X}$ som vi kaller $\bar{X}_{ny}$. Da vil prediksjonsfeilen, $\bar{X}_{ny} - \bar{X}$, ved bruk av $\bar{X}$ som prediktor, være normalfordelt med forventning 0 og standardavvik, $\sigma\sqrt{\frac{2}{5}} = 117000$, og et 95% prediksjonsintervall vil være av formen: predikert verdi $\pm 234\,000$. Denne beregningen indikerer at avviket i dette stratumet nok synes litt ekstremt og muligens tvilsom. De øvrige avvikene derimot virker akseptable ut fra utvalgsstørrelsene.

Forøvrig merker vi oss att de estimerte avdekking-sannsynlighetene i siste kolonne systematisk ligger noe under de observerte relative hyppighetene (jfr. kolonne 4 og 5) der disse er større enn null. Dette er et uttrykk for at vi har kontrollert for den skjevheten som ble introdusert ved å gjøre utvalget på trinn 2 avhengig av screening-variabelen, *Z*, og et uttrykk for at Z har hatt en tilsiktet effekt. Som nevnt i avsnittet etter (8.8), kan disse sannsynlighetene tolkes som prevalens-estimater (for $E_k \bar{p} = E_k\left[P(Y=1 \mid U)\right]$) i de respektive strata.

11. Noen konklusjoner

- Det synes ikke å være noe evidens i data som gir sterk grunn til å opprettholde den kompliserte utvalgsplanen basert på *korrekt rapportering/feilaktig rapportering*. Snarere tvert imot. Det faktum at antall "avdekking" (endringer) blant bedrifter med presumptivt korrekt rapportering synes å være høyere enn tidligere antatt indikerer at rent tilfeldig trekning innenfor hvert region/bransje-stratum antakelig vil være minst like effektiv og har fordelen av å være enklere.
- Det synes fortsatt at *korrekt rapportering/feilaktig rapportering* har en viss effekt i renholdsbransjen, men ikke i de andre to bransjene.
- Av risikofaktorene som inngår i *korrekt rapportering/feilaktig rapportering* er det evidens for at faktoren "Manglende og uriktige oppgaver" har signifikant betydning, mens de andre ikke har det (med forbehold om små datasett for de andre faktorene).
- Det er evidens for at fylke (slått sammen til distrikter) har betydning for sannsynligheten for endring. Det vil si at distriktet Sør (som her består av Vestfold, Buskerud Telemark, Vest-Agder) har signifikant høyere sannsynlighet for avdekking ($Y = 1$) enn de andre distriktene.
- Sannsynligheten for avdekking er signifikant og betydelig lavere for bransjen "godstrafikk på vei" enn for de to andre bransjene.

- Kovariatene *Samlet vurdering* (811) og *MaxAvPoeng* har begge signifikant forklaringskraft for sannsynligheten for endring etter materiell kontroll, og det er grunn til å tro at *Samlet vurdering* (811) har informasjon utover *MaxAvPoeng* for avdekkingsannsynligheten. Dette kan ha betydning for bestemmelse av utvalget på trinn 2 i en eventuell senere undersøkelse. I denne undersøkelsen er kun *MaxAvPoeng* via dummien *Z* benyttet som screening-grunnlag. Et forslag til nytt og potensielt bedre screening-grunnlaag kunne for eksempel være i stedet å definere *Z* = 1 hvis minst en av følgende tre kriterier innavdekkinger: (1) *MaxAvPoeng* > 0,3 *eller* (2) *Samlet vurdering* (811) er 3 eller 4, *eller* (3) faktoren "Manglende og uriktige oppgaver" i *korrekt rapportering/feilaktig rapportering* foreligger.
- Det er to forklaringsvariable som klart skiller seg ut i den forstand at det er sterk evidens for at de er viktige, nemlig *R* (tilstedeværelse av ekstern regnskapsfører) og screenings-variabelen *Z* (funn på trinn 1). *R* synes ikke å ha noen betydning for avdekkingsannsynligheten på trinn 2, men har en klar negativ effekt endringsbeløpet gitt avdekking ($X > 0$). *Z*, på den annen side, har ingen effekt på endringsbeløpet gitt avdekking (endring), men har en klar positiv effekt på sannsynligheten for avdekking og fungerer således som screening.
- Det er klare forskjeller mellom bransjene når det gjelder utbredelse av endringstilfeller med klart lavere tilbøylighet i bransje 60 (godstrafikk på vei) enn de andre to, men ingen vesentlige forskjell i gjennomsnittlig endringsbeløp, muligens med unntak av bransje 51 (engroshandel med klær, sports- og fritidsutstyr) som kan ligge litt høyere (jfr. tabell 10.2). Det er imidlertid uklart hvor mye av denne forskjellen skyldes en modellsvakhet og hvor mye den skyldes en reell forskjell. Det samme forbeholdet gjelder de påfallende høye prevalensanslagene i tabell 10.2 for distrikt Sør.
- Når det gjelder andre distriktsmessige forskjeller, synes det å foreligge et noe lavere gjennomsnittlig endringsbeløp (gitt avdekking) i Øst enn i de andre distriktene (jfr tabell 10.2).
- Ut fra den valgte prediksjonsmodellen har ikke virksomhetens størrelse, målt ved variabelen *Ansatte* (antall ansatte), noen betydning for prevalens-sannsynligheten $P(Y = 1|U)$, eller forventet endringsbeløp gitt endring, $E(X \mid X > 0, U)$. Dette behøver ikke å være helt riktig. Det er nemlig en sterk sammenheng mellom *R* (ekstern regnskapsfører) og *Ansatte* slik at den utelatte variabelen *Ansatte* kan ha indirekte betydning via den inkluderte variabelen *R*. En rekke andre variable som tilsynelatende ikke har betydning som for eksempel virksomhetens struktur (AS, ENK, osv), kommunens næringsstruktur, omsetning osv., kan på den måten ha indirekte betydning via de inkluderte variablene. Det er ofte slik når man velger ut noen forklaringsvariable ut fra en større gruppe av potensielle kovariater at noen av de inkluderte variablene kan fungere delvis som "proxier" for noen av de utelatte. Det er dette som gjør at man ikke uten videre kan tolke signifikante regresjonseffekter (koeffisienter) som kausale effekter.
- Det er betydelig usikkerhet ved total-anslagene i tabell 10.1 som domineres av estimeringsusikkerheten ved estimering av forventet størrelse på endringen. Dette skyldes først og fremst det tynne datagrunnlaget for endringsbeløpenes størrelse (22 observasjoner).

Appendiks 1. Matematiske utledninger

A1.1 Begrunnelse for resultater i (8.8)

Notasjon: For eksempel $E(X \mid U,Y,Z)$, som i utgangspunktet er en funksjon av U,Y,Z, skriver jeg noen ganger som $E_U(X \mid Y,Z)$ for å understreke at jeg ser på funksjonen partielt som en funksjon av Y,Z der U holdes fast, og med fordeling bestemt av den betingete fordelingen for $(Y,Z \mid U)$. For øvrig vil fordelingen under forventning og variansberegninger nedenfor avhenge av stratum *k* som undertrykkes i uttrykkene nedenfor. Argumentasjonen for øvrig bygger på resultater av typen (8.4) - (8.5).

Skriv kort som før $p_z = P(Y=1 \mid U, Z=z)$, $q = P(Z=1 \mid U)$ og $\overline{p} = P(Y=1 \mid U)$. Vi får

$$\overline{p} = E(Y \mid U) = E_U(Y) = E_U\left[E_U(Y \mid Z)\right] =$$
$$= E_U(Y \mid Z=0)P(Z=0 \mid U) + E_U(Y \mid Z=1)P(Z=1 \mid U) = p_0(1-q) + p_1 q$$

som viser (8.8)**i**.

Skriv som før kort $\mu_z = E(X \mid U, Y=1, Z=z) = E_U(X \mid Y=1, Z=z)$. Vi har

$$E(X \mid U) = E_U(X) = E_U\left[E_U(X \mid Y,Z)\right]$$

der

$$E_U(X \mid Y=y, Z=z) = \begin{cases} 0 & \text{for } y=0, \text{ med sannsynlighet } P(Y=0 \mid U) = 1-\overline{p} \\ \mu_0 & \text{for } y=1 \text{ og } z=0, \text{ med sanns.het } (1-q)p_0 \\ \mu_1 & \text{for } y=1 \text{ og } z=1, \text{ med sanns.het } qp_1 \end{cases}$$

Hvis A er en stokastisk variabel som antar verdiene $0, a_1, a_2$ med sannsynligheter r_0, r_1, r_2, h.h.v., er $E(A) = a_1 r_1 + a_2 r_2$ og $\text{var}(A) = a_1^2 r_1 + a_2^2 r_2 - (a_1 r_1 + a_2 r_2)^2$. Av dette følger generelt

(A1) $$E(X \mid U) = E_U\left[E_U(X \mid Y,Z)\right] = \mu_0(1-q)p_0 + \mu_1 qp_1$$

og

(A2) $$\text{var}_U\left[E_U(X \mid Y,Z)\right] = \mu_0^2(1-q)p_0 + \mu_1^2 qp_1 - \left[\mu_0(1-q)p_0 + \mu_1 qp_1\right]^2$$

På den annen side er (fra (8.6))

$$\mathrm{var}_U(X \mid Y=y, Z=z) = \begin{cases} 0 & \text{for } y=0 \text{ med sanns.het } P(Y=0 \mid U) = 1-\overline{p} \\ \varphi\mu_0^2 & \text{for } y=1 \text{ og } z=0, \text{ med sanns.het } (1-q)p_0 \\ \varphi\mu_1^2 & \text{for } y=1 \text{ og } z=1, \text{ med sanns.het } qp_1 \end{cases}$$

som gir

(A3) $$E_U\left[\mathrm{var}_U(X \mid Y,Z)\right] = \varphi\left(\mu_0^2(1-q)p_0 + \mu_1^2 qp_1\right)$$

Dermed

(A4) $$\begin{aligned} \mathrm{var}(X \mid U) &= \mathrm{var}_U\left[E_U(X \mid Y,Z)\right] + E_U\left[\mathrm{var}_U(X \mid Y,Z)\right] = \\ &= (1+\varphi)\left(\mu_0^2(1-q)p_0 + \mu_1^2 qp_1\right) - \left[\mu_0(1-q)p_0 + \mu_1 qp_1\right]^2 \end{aligned}$$

I det spesielle tilfellet at μ_z ikke avhenger av z ($\mu_0 = \mu_1 = \mu$), reduserer (A1) seg til $E(X \mid U) = \mu\overline{p}$ som er (8.8)**ii**, og (A4) reduserer seg til

$$\mathrm{var}(X \mid U) = (1+\varphi)\mu^2\overline{p} - \left(\mu\overline{p}\right)^2 = \varphi\mu^2\overline{p} + \mu^2\overline{p}(1-\overline{p})$$

som er (8.8)**iii**.

Generelt får vi nå

(A5) $$E(X) = E\left[E(X \mid U)\right] = E(\mu_0(1-q)p_0 + \mu_1 qp_1)$$

og

$$\begin{aligned} \mathrm{var}(X) &= E\left[\mathrm{var}(X \mid U)\right] + \mathrm{var}\left[E(X \mid U)\right] = \\ &= (1+\varphi)E\left(\mu_0^2(1-q)p_0 + \mu_1^2 qp_1\right) - E\left[\mu_0(1-q)p_0 + \mu_1 qp_1\right]^2 + E\left[\mu_0(1-q)p_0 + \mu_1 qp_1\right]^2 - \\ &\qquad \left[E(\mu_0(1-q)p_0 + \mu_1 qp_1)\right]^2 \end{aligned}$$

hvorav

(A6) $$\mathrm{var}(X) = (1+\varphi)E\left(\mu_0^2(1-q)p_0 + \mu_1^2 qp_1\right) - \left[E(\mu_0(1-q)p_0 + \mu_1 qp_1)\right]^2$$

Igjen, hvis $\mu_0 = \mu_1 = \mu$, vil (A5) og (A6) redusere seg til (8.8)**iv** og (8.8)**v** h.h.v.

A1.2 Likelihood-funksjonen

Siden *X* (og *Y*) bare er observert for en del av det opprinnelige utvalget, kan denne situasjonen ses på som en situasjon med manglende data ("missing data problem"), som er behørig behandlet i litteraturen. Den manglende del av data er hos oss design-bestemt og avhengig av responsen *Z*. I tillegg er det noen få manglende observasjonsenheter som først og fremst har administrative årsaker som ikke er avhengig av responsvariablene og som vi derfor trygt kan se bort ifra.

La hele datamatrisen (inklusive de manglende data) betegnes med $W = (W_{ij}) = (W_{obs}, W_{mangel})$, der W_{obs}, W_{mangel} angir den delen av *W* som er observert og den delen som mangler. La $f(W;\theta) = f(W_{obs}, W_{mangel};\theta)$ være simultan-tettheten for *W*. Dersom vi ignorerer mangel-mekanismen, er likelihood-funksjonen for våre observerte data er gitt ved

$$(A7) \qquad L(\theta) = f(W_{obs}) = \int f(W_{obs}, W_{mangel}) dW_{mangel}$$

og spørsmålet oppstår om når vi kan ignorere mangel-mekanismen. La $M = (M_{ij})$ være en tilsvarende datamatrise av 0-1-variable der $M_{ij} = 1$ dersom W_{ij} er observert og $M_{ij} = 0$ dersom W_{ij} mangler. Våre data er dermed i realiteten gitt ved (W_{obs}, M) og den egentlige likelihooden er

$$\bar{L}(\theta) = \int f(W_{obs}, W_{mangel}, M;\theta) dW_{mangel} = \int f(W_{obs}, W_{mangel};\theta) f(M \mid W_{obs}, W_{mangel}) dW_{mangel}$$

I vårt tilfelle er mangel-mekanismen gitt ved fordelingen $f(M \mid W_{obs}, W_{mangel})$ som ikke avhenger av θ og som kun avhenger av data via *Z* som inngår i W_{obs}. Dermed må vi ha at $f(M \mid W_{obs}, W_{mangel}) = f(M \mid W_{obs})$, en egenskap som er kalt MAR ("missing at random") i litteraturen og som er en tilstrekkelig betingelse for at vi kan ignorere mangel-mekanismen. For i så fall blir likelihooden

$$\bar{L}(\theta) = f(M \mid W_{obs}) \int f(W_{obs}, W_{mangel};\theta) dW_{mangel} = f(M \mid W_{obs}) L(\theta)$$

og maksimering av $\bar{L}(\theta)$ er ekvivalent med maksimering av $L(\theta)$. Vi kan altså ignorere mangel-mekanismen for våre data og sette opp likelihood-funksjonen i tråd med (A7).

Den fulle datamatrisen (uten manglende observasjoner) er gitt ved $(x_i, y_i, z_i, u_i), \quad i = 1, 2, \ldots, n$, der *n* = 277 med kovariatvektoren *U* definert i (10.1) og som anses for eksogen. Den fulle likelihooden (uten manglende observasjoner) blir da

$$L_{full}(\theta) = \prod_{i=1}^{n} f(x_i, y_i, z_i \mid u_i; \theta)$$

der $f(x_i, y_i, z_i \mid u_i; \theta)$ er simultantettheten for responsen $(X, Y, Z \mid U)$ som er en blanding av kontinuerlig og diskret fordeling. La $q_{z_i} = q_i^{z_i}(1-q_i)^{z_i}$ der $q_i = P(Z=1 \mid U=u_i)$ som før. Dessuten la $p_{z_i} = P(Y=1 \mid U=u_i, Z=z_i)$ Parallelt med utviklingen i **A1.1**, kan vi da skrive

(A8) $$f(x_i, y_i, z_i \mid u_i; \theta) = (1-y_i)q_{z_i}(1-p_{z_i}) + y_i q_{z_i} p_{z_i} g(x_i; \mu_i, \varphi)$$

der *g* er tetthetsfunksjonen i gammafordelingen med parametre (μ_i, φ) og der $\mu_i = E(X \mid U=u_i, Y=1, Z=z_i)$ (som er uavhengig av z_i i den foreslåtte prediksjonsmodellen).

La $I = I_0 \cup I_1$ være mengden av de indekser, *i*, der (x_i, y_i) er observert, $I_0 = \{i \mid y_i = 0\}$ og $I_1 = \{i \mid y_i = 1\}$. Siden fordelingen (q_{z_i}) for $(Z \mid U = u_i)$ opptrer som felles faktor i (A8), får vi likelihooden for de observerte data

$$L = L_{obs} = \int_{\text{manglende data}} L_{full} = \prod_{i \in I} f(x_i, y_i, z_i \mid u_i) \prod_{i \notin I} q_{z_i}$$

$$= \prod_{i=1}^{n} q_{z_i} \prod_{i \in I} \left[(1-p_{z_i})(1-y_i) + y_i p_{z_i} g(x_i; \mu_i, \varphi) \right]$$

som kan skrives

(A9) $$L(\theta) = \prod_{i=1}^{n} q_i^{z_i}(1-q_i)^{z_i} \prod_{i \in I} p_{z_i}^{y_i}(1-p_{z_i})^{1-y_i} \prod_{i \in I_1} g(x_i; \mu_i, \varphi)$$

Vi ser at likelihooden er redusert til et produkt av tre faktorer som avhenger av hvert sitt uavhengige sett av parametre (α, β og γ h.h.v.), som derfor kan maksimeres separat. (φ anses kjent og holdes utenfor maksimeringen. Jfr. A1.3.) I tillegg følger at ML-estimatorene vil være (asymptotisk) uavhengige.

A1.3 Beregning av standardfeil (asymptotisk) for forventet totaltall, $\hat{E}_k(T)$

Som en forenkling holdes dispersjonsparameteren φ utenfor maksimering av likelihooden. Dette er i tråd med vanlig GLM-praksis der likelihooden maksimeres med hensyn på parametrene i den lineære prediktoren, η, mens φ konsentreres ut. Av forskjellige grunner velger man ofte å estimere φ på en annen måte enn ML (for eksempel via såkalte Pearson-residualer). Ved videre beregning av standardfeil anses φ kjent lik estimatet. Dette

forenkler algoritmene betraktelig men fører til at vi ikke får standardfeil for $\hat{\varphi}$. I tillegg til en rekke tekniske grunner begrunnes dette ved å vise til at φ ikke har interesse i seg selv, kun såkalt ansillær (indirekte) interesse. Man velger altså i praksis φ på en eller annen hensiktsmessig måte og studerer kvaliteten av valget ved diagnostisk analyse - for eksempel ved residualer og lignende.

Etter dette lar vi parametervektoren (dimensjon 14) være $\theta' = (\alpha', \beta', \gamma, ')$. La de estimerte kovariansmatrisene for de tre delene være $\Sigma_\alpha, \Sigma_\beta$ og Σ_γ h.h.v. På grunn av faktoriseringen av likelihood-funksjonen blir (den asymptotiske) kovariansmatrisen for $\hat{\theta}$

$$\Sigma_\theta = \text{kovar}(\hat{\theta}) = \begin{pmatrix} \Sigma_\alpha & 0 & 0 \\ 0 & \Sigma_\beta & 0 \\ 0 & 0 & \Sigma_\gamma \end{pmatrix}$$

der 0-ene betegner 0-matriser av passende størrelse. Benytter vi prediktoren i (10.7),

$$\hat{T} = \hat{E}(T) = \sum_{k=1}^{K} \frac{N_k}{n_k} \sum_{i=1}^{n_k} h(U_i; \hat{\theta}) \quad \text{der} \quad h(U_i; \hat{\theta}) = \hat{\mu}_i \hat{\bar{p}}_i,$$

får vi ved første ordens Taylor utvikling

(A10)
$$\hat{E}(T) - E(T) \approx d'(\hat{\theta} - \theta) = d' \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\beta} - \beta \\ \hat{\gamma} - \gamma \end{pmatrix}$$

der

$$d' = \sum_{k=1}^{K} N_k (d_{\alpha k}', d_{\beta k}', d_{\gamma k}')$$

$$d_{\alpha k} = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{\partial}{\partial \alpha} h(U_i, \hat{\theta}) = \frac{1}{n_k} \sum_{i=1}^{n_k} \left[\mu_i (p_{1i} - p_{0i}) q_i (1 - q_i) \right] \cdot \begin{pmatrix} B_{1i} \\ B_{2i} \\ Sør_i \\ Vest_i \\ Ansatte_i \\ 1 \end{pmatrix}$$

$$d_{\beta k} = \frac{1}{n_k}\sum_{i=1}^{n_k}\frac{\partial}{\partial\beta}h(U_i,\hat{\theta}) = \frac{1}{n_k}\sum_{i=1}^{n_k}\left[\mu_i p_{0i}(1-p_{0i})(1-q_i)\right]\cdot\begin{pmatrix} B_{2i} \\ Sør_i \\ 0 \\ 1 \end{pmatrix} + \frac{1}{n_k}\sum_{i=1}^{n_k}\left[\mu_i p_{1i}(1-p_{1i})q_i\right]\cdot\begin{pmatrix} B_{2i} \\ Sør_i \\ 1 \\ 1 \end{pmatrix}$$

$$d_{\gamma k} = \frac{1}{n_k}\sum_{i=1}^{n_k}\frac{\partial}{\partial\gamma}h(U_i,\hat{\theta}) = \frac{1}{n_k}\sum_{i=1}^{n_k}\left[\mu_i \bar{p}_i\right]\cdot\begin{pmatrix} Øst_i \\ Komsentral_i \\ R_i \\ 1 \end{pmatrix}$$

Av (A10) følger nå den asymptotiske standardfeilen for estimatoren $\hat{E}(T)$

(A11) $$\mathrm{SE}(\hat{E}(T)) = \sqrt{d'\Sigma_\theta d}$$

Appendiks 2. Utskrifter

(Beregnet med STATA.)

1 Regresjonsutskrift for modell (1)

```
Logistic regression                                     Number of obs   =         83
                                                        LR chi2(3)      =      14.58
                                                        Prob > chi2     =     0.0022
Log likelihood = -40.705957                             Pseudo R2       =     0.1519

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |  -.0839385   .5532235    -0.15   0.879    -1.168237     1.00036
          B1 |  -.1193765   .6071331    -0.20   0.844    -1.309336    1.070583
          B2 |  -2.452105   .8155751    -3.01   0.003    -4.050603   -.8536077
       _cons |  -.2768205    .484329    -0.57   0.568    -1.226088    .6724468
------------------------------------------------------------------------------
```

2 Regresjonsutskrift med fylke (gruppert til distrikt) inkludert

```
Logistic regression                                     Number of obs   =         83
                                                        LR chi2(7)      =      21.77
                                                        Prob > chi2     =     0.0028
Log likelihood = -37.114484                             Pseudo R2       =     0.2267

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |   .2579286   .6155829     0.42   0.675    -.9485917    1.464449
          B1 |  -.3423358   .6646618    -0.52   0.607    -1.645049    .9603775
          B2 |  -2.803243   .8619945    -3.25   0.001    -4.492721   -1.113765
         Øst |   .4534293   1.300539     0.35   0.727     -2.09558    3.002438
         Sør |    2.01121   1.207318     1.67   0.096    -.3550906     4.37751
        Vest |   .2081132   1.386656     0.15   0.881    -2.509683    2.925909
        Midt |   1.118964   1.284157     0.87   0.384    -1.397938    3.635865
       _cons |  -1.481771   1.191689    -1.24   0.214    -3.817438    .8538958
------------------------------------------------------------------------------
-----
```

3 Regresjonsutskrift med distriktet *Sør* inkludert

```
Logistic regression                                     Number of obs   =         83
                                                        LR chi2(4)      =      20.59
                                                        Prob > chi2     =     0.0004
Log likelihood = -37.704051                             Pseudo R2       =     0.2145

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |   .1296311   .5911831     0.22   0.826    -1.029067    1.288329
          B1 |  -.2587471   .6476504    -0.40   0.690    -1.528118    1.010624
          B2 |  -2.717807   .8568604    -3.17   0.002    -4.397222   -1.038391
         Sør |   1.424616   .6001007     2.37   0.018     .2484397    2.600791
       _cons |  -.8784895   .5735138    -1.53   0.126    -2.002556    .2455769
------------------------------------------------------------------------------
```

4 Regresjonsutskrift med vurderingsvariablene *MaxAvPoeng* og *Samlet vurdering 811* begge inkludert

```
Logistic regression                                     Number of obs   =         83
                                                        LR chi2(5)      =      31.39
                                                        Prob > chi2     =     0.0000
Log likelihood = -32.304571                             Pseudo R2       =     0.3270

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          B1 |  -.3233997   .7807576    -0.41   0.679    -1.853657    1.206857
          B2 |  -2.341527   .9241248    -2.53   0.011    -4.152778   -.5302757
         Sør |    1.79577   .6926837     2.59   0.010      .438135    3.153405
Vurderin~811 |   .8459508   .4453489     1.90   0.057    -.0269171    1.718819
  MaxAvPoeng |    .697331    1.65848     0.42   0.674    -2.553231    3.947893
       _cons |  -3.057871   .9779137    -3.13   0.002    -4.974547   -1.141196
------------------------------------------------------------------------------
```

5 Regresjonsutskrift med bare vurderingsvariabelen *Samlet vurdering 811* inkludert

```
Logistic regression                                     Number of obs   =         83
                                                        LR chi2(4)      =      31.21
                                                        Prob > chi2     =     0.0000
Log likelihood = -32.393274                             Pseudo R2       =     0.3251

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          B1 |  -.4448064   .7261668    -0.61   0.540    -1.868067    .9784544
          B2 |  -2.401978   .9121796    -2.63   0.008    -4.189818    -.614139
         Sør |   1.767561   .6871572     2.57   0.010     .4207577    3.114364
Vurderin~811 |   .9743067   .3285318     2.97   0.003     .3303962    1.618217
       _cons |  -3.126005   .9622329    -3.25   0.001    -5.011947   -1.240064
------------------------------------------------------------------------------
```

6 Regresjonsutskrift med bare vurderingsvariabelen *MaxAvPoeng* inkludert

```
Logistic regression                                     Number of obs   =         83
                                                        LR chi2(4)      =      27.58
                                                        Prob > chi2     =     0.0000
Log likelihood = -34.205552                             Pseudo R2       =     0.2873

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          B1 |    .126899   .7104174     0.18   0.858    -1.265494    1.519291
          B2 |  -2.265814   .8958245    -2.53   0.011    -4.021597     -.51003
         Sør |   1.709189   .6513287     2.62   0.009     .4326085     2.98577
  MaxAvPoeng |   3.017698   1.203743     2.51   0.012     .6584055    5.376992
       _cons |  -1.797957     .64149    -2.80   0.005    -3.055254   -.5406593
------------------------------------------------------------------------------
```

7 Regresjonsutskrift for endelig prediksjonsmodell for avdekkingsannsynligheten i tabell 6.2

```
Generalized linear models                           No. of obs      =         83
Optimization     : ML                               Residual df     =         79
                                                    Scale parameter =          1
Deviance         =  68.23830927                     (1/df) Deviance =  .8637761
Pearson          =  111.6218444                     (1/df) Pearson  =  1.412935

Variance function: V(u) = u*(1-u)                   [Bernoulli]
Link function    : g(u) = ln(u/(1-u))               [Logit]

                                                    AIC             =  .9185338
Log likelihood   = -34.11915463                     BIC             = -280.8501

------------------------------------------------------------------------------
             |                 OIM
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          B2 |  -2.349557   .8569776    -2.74   0.006    -4.029203   -.6699122
           Z |   1.641692   .6244343     2.63   0.009     .4178234    2.865561
         Sør |   1.556261   .6437865     2.42   0.016      .294463     2.81806
       _cons |  -1.657749   .5281197    -3.14   0.002    -2.692845   -.6226534
------------------------------------------------------------------------------
```

8 Regresjonsutskrift for full modell for avdekkingsannsynligheten i tabell 6.2

```
Generalized linear models                         No. of obs      =        83
Optimization     : ML                             Residual df     =        63
                                                  Scale parameter =         1
Deviance         =  56.50025984                   (1/df) Deviance =  .8968295
Pearson          =  116.9536145                   (1/df) Pearson  =  1.856407

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  1.162654
Log likelihood   = -28.25012992                   BIC             = -221.8867

------------------------------------------------------------------------------
             |                 OIM
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |  -.0498875   .7968964    -0.06   0.950    -1.611776    1.512001
          B1 |   .8116752   1.236087     0.66   0.511    -1.611011    3.234361
          B2 |  -1.755352   1.118504    -1.57   0.117    -3.947579    .4368745
           Z |   2.764769   1.018956     2.71   0.007     .7676519    4.761886
 EksternRegn |  -.7788173   1.092221    -0.71   0.476     -2.91953    1.361896
     Ansatte |   .1768366   .1179699     1.50   0.134    -.0543801    .4080533
         Ost |  -2.064412   3.712018    -0.56   0.578    -9.339834     5.21101
         Sor |   .4975556   3.699141     0.13   0.893    -6.752628    7.747739
        Vest |  -1.517542   3.692355    -0.41   0.681    -8.754425    5.719342
        Midt |  -1.370863    3.30983    -0.41   0.679    -7.858012    5.116285
         ENK |   1.518017   2.203699     0.69   0.491    -2.801155    5.837188
          AS |   .2998624   2.164858     0.14   0.890    -3.943181    4.542906
       Alder |  -.0483548   .0600122    -0.81   0.420    -.1659765    .0692669
  komsentral |  -.2887484   .4193303    -0.69   0.491    -1.110621     .533124
_Ikomnaeri~3 |  -.6047876   2.024331    -0.30   0.765    -4.572404    3.362829
_Ikomnaeri~4 |   -25.0145   2232.806    -0.01   0.991    -4401.234    4351.205
_Ikomnaeri~5 |  -1.684211   2.434676    -0.69   0.489    -6.456089    3.087666
_Ikomnaeri~6 |  -1.737414   2.949343    -0.59   0.556     -7.51802    4.043192
_Ikomnaeri~7 |   -.342983   2.438305    -0.14   0.888    -5.121972    4.436006
       _cons |   1.220911    3.47027     0.35   0.725    -5.580695    8.022516
------------------------------------------------------------------------------
```

9 Regresjonsutskrift for redusert full modell for avdekkingsannsynligheten i tabell 6.2

```
Generalized linear models                         No. of obs      =        83
Optimization     : ML                             Residual df     =        68
                                                  Scale parameter =         1
Deviance         =  62.33604689                   (1/df) Deviance =  .9167066
Pearson          =  133.8383974                   (1/df) Pearson  =  1.968212

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  1.112482
Log likelihood   = -31.16802345                   BIC             = -238.1451

------------------------------------------------------------------------------
             |                 OIM
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |  -.0382107   .7185933    -0.05   0.958    -1.446628    1.370206
          B1 |   .0586978   .9197652     0.06   0.949    -1.744009    1.861405
          B2 |  -2.336559   1.037016    -2.25   0.024    -4.369073    -.304045
           Z |    2.16144   .8115346     2.66   0.008     .5708614    3.752018
 EksternRegn |  -.9039271   .8647142    -1.05   0.296    -2.598736    .7908817
     Ansatte |   .0282854   .0394241     0.72   0.473    -.0489843    .1055552
         Ost |  -.0821707   1.655164    -0.05   0.960    -3.326232     3.16189
         Sor |   2.445139   1.576169     1.55   0.121    -.6440951    5.534373
        Vest |   .7047922   1.767236     0.40   0.690    -2.758927    4.168511
        Midt |   1.093009   1.574914     0.69   0.488    -1.993765    4.179784
         ENK |  -.1509667   1.500428    -0.10   0.920    -3.091751    2.789817
          AS |  -.4405202   1.500466    -0.29   0.769    -3.381379    2.500339
       Alder |  -.0248704   .0557927    -0.45   0.656     -.134222    .0844812
  komsentral |  -.1682094   .2682611    -0.63   0.531    -.6939915    .3575727
       _cons |  -.5090621   2.050443    -0.25   0.804    -4.527856    3.509732
------------------------------------------------------------------------------
```

10 Regresjonsutskrift for alternativ prediksjonsmodell for avdekkingsannsynligheten i tabell 6.2

```
Generalized linear models                         No. of obs      =        83
Optimization     : ML                             Residual df     =        78
                                                  Scale parameter =         1
Deviance         =  65.94970167                   (1/df) Deviance =   .845509
Pearson          =  84.22072683                   (1/df) Pearson  =  1.079753

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .9150566
Log likelihood   = -32.97485084                   BIC             = -278.7199

------------------------------------------------------------------------------
             |                 OIM
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          B2 |   -2.47607   .8702819    -2.85   0.004    -4.181792   -.7703492
           Z |   1.900478   .6809494     2.79   0.005     .5658417    3.235114
         Sor |   2.065917   .7708922     2.68   0.007      .554996    3.576838
        Midt |   1.356039   .9040935     1.50   0.134    -.4159518     3.12803
       _cons |  -2.212449   .6939725    -3.19   0.001     -3.57261   -.8522878
------------------------------------------------------------------------------
```

11 Regresjonsutskrift for prediksjonsmodell pluss tilleggsvariable i tabell 6.3 (redusert datasett)

```
Generalized linear models                         No. of obs      =        61
Optimization     : ML                             Residual df     =        49
                                                  Scale parameter =         1
Deviance         =  43.53875499                   (1/df) Deviance =   .888546
Pearson          =  73.50113646                   (1/df) Pearson  =  1.500023

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  1.107193
Log likelihood   = -21.76937749                   BIC             = -157.8941
```

Y	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
I	.042132	.9078523	0.05	0.963	-1.737226	1.82149
B2	-3.043382	1.112403	-2.74	0.006	-5.223651	-.8631132
Z	1.549373	1.048674	1.48	0.140	-.5059909	3.604738
Sor	2.568652	1.069981	2.40	0.016	.4715265	4.665777
Midt	1.594265	1.334651	1.19	0.232	-1.021603	4.210133
Driftsinn~05	-3.13e-07	7.41e-07	-0.42	0.672	-1.76e-06	1.14e-06
Leielokal~05	-1.40e-06	7.12e-06	-0.20	0.845	-.0000153	.0000126
Driftsres~05	2.39e-06	5.42e-06	0.44	0.660	-8.24e-06	.000013
Næringsin~05	-1.45e-06	5.77e-06	-0.25	0.802	-.0000128	9.86e-06
oms2005	5.25e-07	6.59e-07	0.80	0.426	-7.66e-07	1.82e-06
oms2006	-1.11e-07	3.99e-07	-0.28	0.782	-8.93e-07	6.72e-07
_cons	-2.634694	1.213789	-2.17	0.030	-5.013677	-.255712

12 Regresjonsutskrift for prediksjonsmodell i tabell 6.3 (redusert datasett)

```
Generalized linear models                         No. of obs      =        61
Optimization     : ML                             Residual df     =        57
                                                  Scale parameter =         1
Deviance         =  48.64557079                   (1/df) Deviance =  .8534311
Pearson          =   80.4778079                   (1/df) Pearson  =  1.411891

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .9286159
Log likelihood   = -24.32278539                   BIC             = -185.6742
```

Y	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
B2	-2.287274	.8841552	-2.59	0.010	-4.020186	-.5543613
Z	.9191133	.7490681	1.23	0.220	-.5490331	2.38726
Sor	1.82343	.7400009	2.46	0.014	.3730546	3.273805
_cons	-1.506444	.6369206	-2.37	0.018	-2.754785	-.2581027

13 Regresjonsutskrift for prediksjonsmodell pluss *feilaktig rapportering* i tabell 6.3 (fullt datasett)

```
Generalized linear models                         No. of obs      =        83
Optimization     : ML                             Residual df     =        78
                                                  Scale parameter =         1
Deviance         =  68.11485138                   (1/df) Deviance =  .8732673
Pearson          =  116.4039585                   (1/df) Pearson  =  1.492358

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .9411428
Log likelihood   = -34.05742569                   BIC             = -276.5547

------------------------------------------------------------------------------
             |                 OIM
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |  -.2226125   .6343195    -0.35   0.726    -1.465856    1.020631
          B2 |  -2.352375   .8549881    -2.75   0.006    -4.028121   -.6766294
           Z |   1.689839   .6432617     2.63   0.009     .4290696    2.950609
         Sor |   1.534327     .64697     2.37   0.018     .2662887    2.802365
       _cons |   -1.53883   .6245072    -2.46   0.014    -2.762842   -.3148189
------------------------------------------------------------------------------
```

14 Regresjonsutskrift for full modell for forventet endringstall gitt avdekking

```
Generalized linear models                         No. of obs      =        22
Optimization     : ML                             Residual df     =         7
                                                  Scale parameter =  1.324777
Deviance         =  18.26410739                   (1/df) Deviance =  2.609158
Pearson          =  9.273439845                   (1/df) Pearson  =  1.324777

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

                                                  AIC             =  27.30673
Log likelihood   = -285.3740053                   BIC             =  -3.37319

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          B1 |  -.5131972   1.402986    -0.37   0.715    -3.262999    2.236604
          B2 |   .9195413   1.493893     0.62   0.538    -2.008435    3.847518
           Z |     .46361   2.123572     0.22   0.827    -3.698515    4.625735
           M |   .3270079   2.896075     0.11   0.910    -5.349195    6.003211
         Ost |  -2.462447   1.602797    -1.54   0.124    -5.603872    .6789779
         Sor |  -.8890168   2.068349    -0.43   0.667    -4.942906    3.164872
        Vest |  -1.477969   2.183825    -0.68   0.499    -5.758187    2.802249
          AS |   .5152351   1.005366     0.51   0.608    -1.455247    2.485717
           R |  -1.961653   .9540491    -2.06   0.040    -3.831555   -.0917508
           I |    .181894   1.205909     0.15   0.880    -2.181644    2.545432
  komsentral |   .3653082   .3776526     0.97   0.333    -.3748773    1.105494
    Omsetn06 |  -3.56e-08   1.41e-07    -0.25   0.800    -3.12e-07    2.41e-07
       alder |  -.0650962   .0885012    -0.74   0.462    -.2385555     .108363
    Industri |   .9293011   1.184342     0.78   0.433    -1.391966    3.250568
       _cons |   11.73367   2.290419     5.12   0.000     7.244528     16.2228
------------------------------------------------------------------------------
```

15 Regresjonsutskrift for redusert modell for forventet endringstall gitt avdekking

```
Generalized linear models                         No. of obs      =        22
Optimization     : ML                             Residual df     =        18
                                                  Scale parameter =   .691649
Deviance         =  23.76975174                   (1/df) Deviance =  1.320542
Pearson          =  12.44968195                   (1/df) Pearson  =   .691649

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

                                                  AIC             =  26.55698
Log likelihood   = -288.1268275                   BIC             = -31.86901

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           R |  -1.779677   .5136796    -3.46   0.001     -2.78647   -.7728833
  komsentral |   .3337667   .1190329     2.80   0.005     .1004665     .567067
         Ost |  -1.753912   .6388386    -2.75   0.006    -3.006013   -.5018115
       _cons |    11.6579    .613345    19.01   0.000     10.45576    12.86003
------------------------------------------------------------------------------
```

16 Regresjonsutskrift for full modell for funn på trinn 1 (screening)

```
Generalized linear models                         No. of obs      =       276
Optimization     : ML                             Residual df     =       256
                                                  Scale parameter =         1
Deviance         =  145.0179019                   (1/df) Deviance =  .5664762
Pearson          =  291.2776923                   (1/df) Pearson  =  1.137803

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .6703547
Log likelihood   = -72.50895095                   BIC             = -1293.805

------------------------------------------------------------------------------
             |                 OIM
           Z |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           I |  -.5816859   .4550234    -1.28   0.201    -1.473515    .3101436
          B1 |  -1.092893   .6272067    -1.74   0.081    -2.322196    .1364091
          B2 |  -2.016995   .6055183    -3.33   0.001    -3.203789   -.8302005
           R |   .4874139   .6257377     0.78   0.436    -.7390095    1.713837
     Ansatte |  -.2163971   .1223795    -1.77   0.077    -.4562565    .0234622
         Ost |   .1282347   .7017905     0.18   0.855    -1.247249    1.503719
         Sor |  -2.398583   .8655814    -2.77   0.006    -4.095091   -.7020742
        Vest |  -1.717654   .8627861    -1.99   0.047    -3.408683   -.0266242
        Midt |   -1.34587   .7856899    -1.71   0.087    -2.885794     .194054
         ENK |   .4575708   1.037653     0.44   0.659    -1.576192    2.491334
          AS |  -.2921491   1.050341    -0.28   0.781     -2.35078    1.766482
       Alder |  -.0529451   .0307142    -1.72   0.085    -.1131439    .0072537
  komsentral |   .1139704   .1990086     0.57   0.567    -.2760793      .50402
    komnaer2 |  -2.584328   1.871325    -1.38   0.167    -6.252058    1.083403
    komnaer3 |  -1.357666   1.533534    -0.89   0.376    -4.363337    1.648005
    komnaer4 |  -16.80895   1046.899    -0.02   0.987    -2068.694    2035.076
    komnaer5 |  -1.858577   1.659087    -1.12   0.263    -5.110328    1.393173
    komnaer6 |  -1.682334   1.598295    -1.05   0.293    -4.814935    1.450268
    komnaer7 |  -1.670222   1.706298    -0.98   0.328    -5.014504     1.67406
       _cons |    1.48254   1.606475     0.92   0.356    -1.666092    4.631173
------------------------------------------------------------------------------
```

17 Regresjonsutskrift for redusert modell (prediksjonsmodell) for funn på trinn 1

```
Generalized linear models                         No. of obs      =       277
Optimization     : ML                             Residual df     =       271
                                                  Scale parameter =         1
Deviance         =  161.3396896                   (1/df) Deviance =  .5953494
Pearson          =  303.1812066                   (1/df) Pearson  =   1.11875

Variance function: V(u) = u*(1-u)                 [Bernoulli]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .6257751
Log likelihood   = -80.66984479                   BIC             = -1362.769
```

Z	OIM Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
B1	-1.199268	.5005088	-2.40	0.017	-2.180247	-.2182887
B2	-2.083633	.5454974	-3.82	0.000	-3.152788	-1.014478
Sor	-1.537591	.5845969	-2.63	0.009	-2.68338	-.3918026
Vest	-1.266519	.6713374	-1.89	0.059	-2.582316	.0492785
Ansatte	-.2793843	.1055032	-2.65	0.008	-.4861668	-.0726019
_cons	-.2035987	.3447203	-0.59	0.555	-.8792381	.4720408

18 Oversikt over utvalgene på trinn 1 og trinn 2.

Bransje	Fylke	Antall avdekking på trinn 2		Utvalgsstørrelser trinn 2		Utvalgsstørrelser trinn 1	
		Fullst.	Mangelf	Fullst.	Mangelf.	Fullst.	Mangelf
51	Østfold			2	1	2	6
Engroshandel	Akershus					4	6
Med klær, sports-	Buskerud					2	5
Og fritidsutstyr mv.	Vestfold	1	3	1	3	1	8
	Telemark					3	2
	Vest-Agder	1		1	2	4	3
	Rogaland				2	2	9
	Hordaland					1	
	Møre og Romsdal	2		2		4	1
	Sør-Trøndelag				2	2	4
	Nordland			1	1	2	3
51 Total		4	3	6	12	27	47
60	Østfold			3	1	4	8
Godstransport	Akershus			1	3	2	7
På vei	Buskerud			1	1	3	6
	Vestfold			1	1	3	8
	Telemark	1		3	2	5	9
	Vest-Agder			3	1	5	7
	Rogaland				1	2	7
	Hordaland				1	1	6
	Møre og Romsdal			2	1	4	8
	Sør-Trøndelag		1	1	4	3	10
	Nordland			1	2	3	9
60 Sum		1	1	16	18	35	85
74	Østfold			2		4	4
Rengjøring	Akershus		3		5	2	5
	Buskerud			1	1	3	7
	Vestfold		2	1	2	2	6
	Telemark	1	1	2	1	6	3
	Vest-Agder	2		2		4	4
	Rogaland		1		2	2	6
	Hordaland		1	1	3	2	8
	Møre og Romsdal	1		2	2	2	8
	Sør-Trøndelag			1		3	5
	Nordland		1	1	2	2	9
74 Sum		4	9	13	18	32	65
Total sum		22		83		291	

Appendix G

Harald Goldstein
Supplementær statistisk analyse – data fra 2005 og 2006
Revidert september 2008

Rapport 2: Supplementær statistisk analyse - data fra 2006 og 2005.

- **Analyse av endringer av typen "nettoinntekt bortsett fra feilperiodiseringer og feil bruk av mva-satser" - og litt om sannsynligheten for avdekking av typen "*endringer - alle typer*".**

1. Innledning

Dette er en supplementær rapport basert på forslag som kom fram under høringen av hovedrapporten for statistisk analyse (Rapport 1: Statistisk analyse – data fra 2006, jfr vedlegg H i hovedrapporten). Referanser til hovedrapporten for statistisk analyse betegnes med HRS nedenfor.

Antall endringer av typen "sum-aarsak" i 2006-dataene var bare 19. I håp om å bedre informasjonsgrunnlaget når det gjelder denne typen av endringer ble det foreslått å kombinere 2006-dataene (innhentet 2007) med informasjon fra pilotdataene fra 2004/2005 (innhentet i 2006). Pilotdataene vil bli betegnet med "2005-dataene" nedenfor.

For at denne sammenslåingen av data skal være meningsfull, må to viktige forbehold tas:

- Det kan ha skjedd strukturelle og konjunkturmessige endringer fra 2005 til 2006. Det er innført en årsdummy, kalt *T* nedenfor, for å kontrollere for slike endringer. Det viser seg at *T* har betydning i flere av analysene som tyder på at det har vært endringer. En full analyse av dette spørsmålet, imidlertid, krever i tillegg en analyse av mulige samspill mellom *T* og andre forklaringsvariable - som ikke er forsøkt her siden datagrunnlaget antakelig er for tynt til å få noe meningsfylt ut om samspill. Det første forbehold er derfor at *T* uten samspill klarer å fange opp det vesentlige av relevante strukturelle og konjunkturmessige endringer.
- Det kan ha skjedd en endring av praksis fra 2006 til 2007 når det gjelder gjennomføringen av kontrollene på trinn 1 blant annet som følge av erfaringene gjort under pilotundersøkelsen innhentet i 2006. Den foreliggende analysen finner ingen evidens for at det har skjedd noen vesentlig endring i praksis når det gjelder kriteriet kalt *MaksAvPoeng*, men det kan ha skjedd en endring i forbindelse med kriteriet kalt *samlet vurdering* (811) (jfr. avsnitt 5). Det andre forbeholdet er dermed at det ikke har skjedd noen endring av praksis i kontrollene på trinn 1 spesielt når det gjelder bruken av kriteriet *samlet vurdering* (811).

Siden kun bransjene 51 -*Engroshandel med klær, sports- og fritidsutstyr mv* (betegnet kort med "*Engros*" nedenfor[12]) og 74 - *Rengjøring*, er felles for 2005- og 2006-dataene, vil denne supplerende analysen konsentrere seg om disse to bransjene - som uansett omfatter de fleste avdekkingene. De ikke-overlappende bransjene var "databehandling" (2005) og "60 - godstrafikk på vei" (2006).

Dette økte antall endringer av typen "sum-aarsak" til ialt 24.

Det ble besluttet å gjennomføre den supplementære analysen i to versjoner:

Alternativ 1 - Observasjonene for *En gros* og *Rengjøring* fra 2005 og alle 2006-observasjonene benyttes.

Alternativ 2 - Kun observasjoner for bransjene *En gros* og *Rengjøring* benyttes

Tabell 1.1 Oversikt over antall observasjoner for sammenslått materiale (Avdekking-1 betyr avdekking av typen "sum-aarsak")

Data innhentet		Trinn 1	Trinn 2	Avdekking-1
2007	Alternativ 1	291	83	19
	Alternativ 2	171	49	18
2006	Alternativ 1	128	54	5
	Alternativ 2	128	54	5
I alt	Alternativ 1	419	137	24
	Alternativ 2	299	103	23

Når det gjelder fylker representert i data er det bare delvis overlapping:

Tabell 1.2 Oversikt over fylker virksomheter er trukket fra

2005 data (pilot)	Østfold, Akershus, Hedmark, Oppland, Buskerud, Vestfold, Telemark
2006 data	Østfold, Akershus, Buskerud, Vestfold, Telemark, Vest-Agder, Rogaland, Hordaland, Møre og Romsdal, Sør-Trøndelag og Nordland

Til tross for at HRS tyder på visse regionale forskjeller både i avdekking-sannsynligheter og gjennomsnittlige endringsstørrelser, er, blant annet på grunn av dette, distrikt utelatt fra forklaringsvariablene i denne rapporten. Dette innebærer at estimerte avdekking-sannsynligheter og gjennomsnittlige endringsstørrelser i denne analysen vil kunne tolkes som gjennomsnitt over eventuelle regionale forskjeller. Modellteknisk innebærer dette: Utvalget er fortsatt å anse som stratifisert over *fylke* og *presumptivt feilaktig rapportering* (se HR) samt avhengig på trinn 2 av

[12] Skrives egentlig *en gros* på "norsk", men jeg har tatt meg friheten å fjerne mellomrommet siden uttrykket brukes som variabelnavn.

funn på trinn 1 (via $Z = Z_{mav}$ definert nedenfor), men de betingete fordelingene for avdekking og endringsbeløpets størrelse, gitt eksogene forklaringsvariable, postuleres felles over strata.

I denne studien vil altså fokus dreies litt fra blant annet å måle bransjemessige og regionale forskjeller til mer å finne betydningsfulle faktorer ved virksomhetene innenfor en gruppe av bransjer (her *Engros* og *Rengjøring*) og uavhengig av region. De potensielt betydningsfulle faktorene er karakterisert ved et sett av dikotome variable som definert i avsnitt 2.3.

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 HRS og begrunnet der. Begrunnelsen vil derfor ikke bli gjentatt i denne rapporten.

2. Variable

Framstillingen under er relativt knapp når det gjelder nummerisk beskrivelse av variablene siden mer utførlige deskriptive beskrivelser er utarbeidet av SKD annet sted (se P.G. Larsen, vedlegg G i hovedrapporten). En del deskriptiv informasjon er imidlertid innarbeidet i teksten på steder der informasjonen er relevant.
.

2.1 Responsvariable trinn 1

Vi skiller mellom to typer funn på trinn 1. Begge kan uttrykkes ved dikotome 0-1-variable (dummier):

$$Z_{mav} = \begin{cases} 1 & \text{hvis } \textit{vurderingsskåre}(MAV) > 0{,}3 \\ 0 & \text{hvis } \textit{vurderingsskåre}(MAV) \le 0{,}3 \end{cases}$$

der *vurderingsskåre* er en skåre på skala fra 0 til 1, beregnet på grunnlag av revisors vurdering på trinn 1 av en rekke forhold. Kalt *MaxAvPoeng* eller *MAV* i utskrifter. Z_{mav} er det samme som *Z* i hovedrapporten og utgjør screening-variabelen som ble benyttet til å effektivisere utvalget (øke avdekkingssannsynlighetene) på trinn 2.

$$Z_{sv} = \begin{cases} 1 & \text{hvis } \textit{samlet vurdering}\,(811) \text{ får verdi 3 eller 4} \\ 0 & \text{ellers} \end{cases}$$

I tillegg innføres en alternativ dummy for "*samlet vurdering* (811)", kalt Z_{altsv}, som viser seg nyttig nedenfor i avsnitt 4 under karakteriseringen av fordelingen for endringsbeløpets størrelse gitt avdekking, samt for analysen av i hvilken grad *samlet vurdering* (811) har informasjon om avdekkingssannsynligheten på trinn 2 utover kriteriet MAV > 0.3 (avsnitt 3.6).

$$Z_{altsv} = (1 - Z_{mav}) \cdot Z_{sv} = \begin{cases} 1 & \text{hvis } Z_{mav} = 0 \text{ og } Z_{sv} = 1 \\ 0 & \text{ellers} \end{cases}$$

2.2 Responsvariable trinn 2

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

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

X og Y opptrer i to versjoner (se P.G. Larsens notat "Datamateriale - Random audit " fra desember 2007). Ingen av versjonene omfatter feilperiodiseringer:

- X_1 ("sum-aarsak") omfatter endringer som er knyttet til endringer i nettoinntekt, men ikke endringer knyttet til feil bruk av mva-satser. Y_1 er en tilsvarende avdekkings-indikator (= 1 hvis $X_1 > 0$ og = 0 ellers).
- X_2 ("endringer - alle typer") omfatter det samme som X_1 pluss endringer knyttet til feil bruk av mva-satser, med Y_2 som tilsvarende avdekkings-indikator.
- Denne rapporten vil hovedsakelig fokusere på X_1 og Y_1 (med unntak av avsnitt 3.4 og 3.5).
- X_2 og Y_2 - "endringer - alle typer" ble diskutert i hovedrapporten (HRS) der de ble kalt X og Y henholdsvis.

Avdekking av typen "endringer - alle typer" ($Y = Y_2 = 1$) betegnes med uttrykket "*avdekking*".

Avdekking av typen "sum-aarsak" ($Y_1 = 1$) betegnes nedenfor med uttrykket "*avdekking*-1".

Blant de 83 virksomhetene trukket ut på trinn 2 for 2006-dataene var det 19 *avdekking*-1 og 22 avdekking. Slått sammen med dataene fra pilotundersøkelsen ble det til sammen 23 avdekking-1 og 29 avdekking.

2.3 Forklaringsvariable (såkalte "andre kovariater")

$$T = \text{År} = \begin{cases} 1 & \text{for data innhentet 2007 (2006 -data)} \\ 0 & \text{for data innhentet 2005 (dvs. pilot-dataene fra 2004)} \end{cases}$$

Ansatte0 - Dummy = 1 for null antall ansatte og = 0 ellers.

R - Dummy = 1 for ekstern regnskapsfører og = 0 ellers.

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

Engros - Dummy = 1 for bransje: 51 -Engroshandel med klær, sports- og fritidsutstyr mv., og = 0 ellers.

Rengj - Dummy =1 for bransje: 74 - Rengjøring, og = 0 ellers.

EngrosRengj - Dummy =1 for bransje: Engros eller Rengjøring, og = 0 ellers.

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

$$Komsentral = \begin{cases} 1 & \text{hvis kommunen mest sentral (jfr. SSB definisjon 7 = mest sentral)} \\ 0 & \text{ellers (dvs. 1 - 6 ifølge SSB definisjon)} \end{cases}$$

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

$$KSminKTJ = Komsentral - Komtjenest$$

Merknader

- Antall ansatte har blitt erstattet av dummien *Ansatte0*. 47% av virksomhetene i materialet hadde 0 ansatte, og det viste seg at *Ansatte0* stort sett ga noe bedre tilpasning enn *Ansatte* i regresjonsmodellene som ble prøvet.
- Variablene ENK og *Ansatte*0 har tilstrekkelig uavhengig variasjon til at begge kan være med regresjonsanalysene, som vist i tabell 2.1.

Tabell 2.1 Frekvenstabell for ENK og *Ansatte*0

	*Ansatte*0		
ENK	0	1	Sum
0	167	54	221
1	51	147	198
Sum	218	201	419

- Imidlertid viste det seg til slutt at *Ansatte*0 ikke oppnådde signifikans i noen av de foreslåtte prediksjonsmodellene nedenfor, eller nær signifikans i full-modellene, og tolkes dermed å ikke ha vesentlig betydning for simultanfordelingen for responsvariablene. Det kan synes som en eventuell effekt av *Ansatte*0 blir dominert av effekten av variablene ENK og *Nyreg*, som blant annet kommer til syne i fordelingen for funn på trinn 1.
- Alle forklaringsvariable er dikotome i denne analysen bortsett fra *KSminKTJ* som tar tre verdier, 1, 0 og -1.
- En konklusjon i HRS var at variabelen revisors *samlet vurdering* (811) hadde informasjon utover skår-variabelen *MaksAvPoeng* for sannsynligheten for avdekking på trinn 2, slik at begge variablene burde inngå i screeningsvariabelen ved en eventuell senere undersøkelse. Derfor vil begge indikatorene Z_{sv} og Z_{mav} inngå i analysen under i motsetning til HRS som bare brukte Z_{mav} i regresjonsanalysen.
- På grunn av Z_{mav}'s innflytelse på trekkingen av trinn-2-utvalget bør Z_{mav} prioriteres i en prediksjons-modell der begge variablene hver for seg synes å ha betydning men ikke samlet. Dette viser seg aktuelt for analysen av sannsynligheten for avdekking-1. Innebyrdes fordeling i materialet mellom Z_{sv} og Z_{mav} er vist i tabell 2.2. Det lave totaltallet 401 (av totalt 419) skyldes i alt 18 manglende observasjoner for Z_{mav} (hvorav 6 stammer fra bransjene *Engros* og *Rengjøring*).

Tabell 2.2 Frekvenstabell for Z_{sv} og Z_{mav}, sammenslått materiale

Samlet Vurdering (811)	*MaksAvPoeng*		
	$Z_{mav}=0$	$Z_{mav}=1$	Sum
$Z_{sv}=0$	333	5	338
$Z_{sv}=1$	43	20	63
Sum	376	25	401

I tabell 2.3-2.5 nedenfor følger en oversikt over variablene i analysen.

Tabell 2.3 Frekvenstabell for dikotome variabler i analysen. Datagrunnlag alternativ 1.

TRINN1	Zmav	Zsv	Zaltsv	T	Engros	Rengj	EngrosRengj
Totalt antall	401	419	401	419	419	419	419
Antall = 1	25	63	43	291	107	192	299
%	6	15	11	69	26	46	71
	Nyreg	ENK	Ansatte0	R	Komsentral	KomTjenest	
Totalt antall	419	419	419	409	419	419	
Antall = 1	93	198	201	307	203	130	
%	22	47	48	75	48	31	
TRINN 2	Zmav	Zsv	Zaltsv	T	Engros	Rengj	EngrosRengj
Totalt antall	135	137	135	137	137	137	137
Antall = 1	20	35	18	83	31	72	103
%	15	26	13	61	23	53	75
	Nyreg	ENK	Ansatte0	R	Komsentral	KomTjenest	
Totalt antall	137	137	137	137	137	137	
Antall = 1	28	70	69	105	67	45	
%	20	51	50	77	49	33	

Tabell 2.4 Frekvenstabell for KSminKTJ = Komsentral – Komtjenest. Datagrunnlag alternativ 1.
(Ks = Komsentral, Ktj = Komtjenest)

TRINN 1	Ks=0, Ktj=1	Ks=Ktj	Ks=1, Ktj=0	Sum
Antall	37	272	110	419
%	9	65	26	100
TRINN 2	Ks=0, Ktj=1	Ks=Ktj	Ks=1, Ktj=0	Sum
Antall	11	93	33	137
%	8	68	24	100

Tabell 2.5 Oversikt over endringstall (*X*) av typen "sum-aarsak". Datagrunnlag alternativ 1. Kun trinn 2.

	Antall	Gj.snitt	St.avvik	Min	Maks
X	23	205 534	249 374	4 000	1 200 000

3. Betydning av andre kovariater og funn på trinn 1 for sannsynligheten for avdekking-1 og avdekking på trinn 2

3.1 Sannsynligheten for avdekking-1 (av typen ”sum-aarsak”). Datagrunnlag alternativ 1.

Tilsvarende framgangsmåte som i avsnitt 6 i HRS ga følgende tabell over regresjonsresultatene for avdekking-1, full modell og to alternative reduserte prediksjonsmodeller. I prediksjonsmodell 2 er bransjene *Engros* og *Rengjøring* slått sammen til en felles indikator kalt *EngrosRengj*.

Tabell 3.1a Regresjonsresultater (logistisk regresjon) for avdekking av typen “sum-aarsak”. Sammenslått materiale. Datagrunnlag alternativ 1.

(Basert på utskrift Ut 1-3 i appendiks A2.1)

Avhengig Y_1	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
T	1.5675	0.013	1.7434	0.003	1.7537	0.003
Zmav	0.3669	0.672	1.1910	0.044	1.1733	0.047
Zsv	0.6205	0.386	----	----	----	----
Engros	2.5580	0.030	2.8963	0.011	----	----
Rengj	2.3669	0.036	2.7591	0.013	----	----
EngrosRengj	----	----	----	----	2.8130	0.010
Nyreg	0.8274	0.190	----	----	----	----
ENK	0.6446	0.358	----	----	----	----
Ansatte0	0.1981	0.760	----	----	----	----
R	-0.5859	0.345	----	----	----	----
Komsentral	-0.8105	0.186	-0.8574	0.140	-0.8679	0.135
KomTjenest	1.1736	0.059	0.9764	0.091	0.9814	0.090
KSminKTJ	----	----	----	----	----	----
konstant	-5.2940	0.000	-5.3313	0.000	-5.3372	0.000
Antall obs	135		135		135	
Log-likelihood	-46.4236		-49.2047		-49.2351	
-2 log LR			5.5622		5.623	
P-verdi redusert vs full modell			0.474		0.467	

Vi ser at årsdummien, *T*, begge bransjeindikatorene samt tjenesteytende kommune er signifikante i den fulle modellen, så disse bør helst være med i prediksjonsmodellen. I den fulle modellen er Z_{sv} og Z_{mav} ikke signifikante, enten de opptrer sammen som i tabell 3.1, eller de opptrer enkeltvis. Hvis Z_{mav} opptrer alene i den fulle modellen får den en p-verdi på 0.20, og Z_{sv} alene en p-verdi på 0.13.

Z_{sv} og Z_{mav}, tatt sammen i prediksjonsmodellene, viser seg ikke å være signifikante. Hver for seg, imidlertid, viser de seg signifikante eller nær signifikante. Z_{mav} er da gjennomgående foretrukket i prediksjonsmodellen siden den influerer på trekningen av trinn-2 utvalget.

Bransjene Engros og Rengjøring er klart signifikante, men effekten er svært lik for begge. Bransjene er derfor slått sammen til en felles indikator i prediksjonsmodell 2 (PM2) som synes å gi en tanke bedre tilpasning enn PM1.

Effekten av *Komsentral* og *Komtjenest* oppnår ikke full signifikans, men i nærheten. Verdiene av effektene er imidlertid ganske like med motsatt fortegn. Det er derfor rimelig å prøve og postulere at de teoretiske regresjonskoeffisientene er like i absolutt verdi men med motsatt fortegn. Dette er ekvivalent med å erstatte de to variablene med *KSminKTJ = Komsentral – Komtjenest* og motiverer PM3 i tabell 3.1b, som gir en klart forbedret tilpasning.

Til tross for forbedringen er ikke signifikansen av Z_{mav} eller *KSminKTJ* spesielt sterk, noe som motiverer PM4 i tabell 3.1b.

Tabell 3.1b Regresjonsresultater (logistisk regresjon) for avdekking av typen "sum-aarsak". Sammenslått materiale. Datagrunnlag alternativ 1.

(Basert på utskrift Ut 4-5 i appendiks A2.1)

Avhengig Y_1	Full modell		Prediksjonsmodell 3		Prediksjonsmodell 4	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
T	1.5675	0.013	1.7667	0.002	1.7388	0.002
Zmav	0.3669	0.672	1.1857	0.043	----	----
Zsv	0.6205	0.386	----	----	----	----
Engros	2.5580	0.030	----	----	----	----
Rengj	2.3669	0.036	----	----	----	----
EngrosRengj	----	----	2.8048	0.010	2.9529	0.005
Nyreg	0.8274	0.190	----	----	----	----
ENK	0.6446	0.358	----	----	----	----
Ansatte0	0.1981	0.760	----	----	----	----
R	-0.5859	0.345	----	----	----	----
Komsentral	-0.8105	0.186	----	----	----	----
KomTjenest	1.1736	0.059	----	----	----	----
KSminKTJ	----	----	-0.9251	0.059	----	----
konstant	-5.2940	0.000	-5.2964	0.000	-5.2353	0.000
Antall obs	135		135		137[13]	
Log-likelihood	-46.4236		-49.2518		-53.3892	
-2 log LR			5.6564		13.9312	
P-verdi redusert vs full modell			0.580		0.125	

[13] Merk at Z_{mav} har to manglende observasjoner som kommer i tillegg når Z_{mav} ikke er med i regresjonen.

Både PM3 og PM4 synes aktuelle som prediksjonsmodeller. Ingen av dem kan forkastes relativt til den fulle modellen. PM4 har fordelen av å være enklere, mens PM3 sier mer om potensielle effekter. Informasjonskriteriene AIC og BIC (ikke rapportert her) har forskjellig preferanse. BIC peker på PM4 som klar "vinner" men AIC holder en svak knapp på PM3.

Det er interessant at *Komsentral* og *Komtjenest* først og fremst synes å ha betydning der de har forskjellig verdi - altså (en negativ effekt på avdekking-1-sannsynligheten) for sentrale kommuner som ikke først og fremst er tjenesteytende og (en positiv effekt) for ikke-sentrale kommuner som først og fremst er tjenesteytende. Innebyrdes fordeling av *Komsentral* og *Komtjenest* i materialet er gitt i tabell 3.2.

Tabell 3.2 Innebyrdes fordeling av *Komsentral* og *Komtjenest*

	Komsentral		
Komtjenest	0	1	Sum
0	179	110	289
1	37	93	130
Sum	216	203	419

Variablene ENK, *Ansatte*0 og *R* (ekstern regnskapsfører) falt klart ut under utviklingen av prediksjonsmodellene og synes ikke å ha vesentlig direkte betydning for avdekking-1-sannsynligheten. Merk at dette ikke nødvendigvis betyr at de ikke betyr noe. For eksempel, ENK, viser seg nedenfor å influere på fordelingen for Z_{mav}, og derigjennom indirekte på avdekking-1-sannsynligheten om PM3 legges til grunn.

Om vi legger til Z_{sv} i PM3, blir verken Z_{mav} eller Z_{sv} signifikante. Mao., når det gjelder sannsynligheten for avdekking-1, synes ikke Z_{sv} å ha informasjon utover Z_{mav}. Z_{mav} er valgt til å være med siden denne er med på bestemme hvordan utvalget på trinn 2 er trukket. Spørsmålet om Z_{sv} har informasjon utover Z_{mav} for avdekkings-sannsynligheten vil bli drøftet nærmere i avsnitt 3.5.

Tabell 3.3 Estimerte felles avdekking-1-sannsynligheter, betinget av Z_{mav}, for bransjene *Engros* og *Rengjøring*. Basert på prediksjonsmodell PM3 og PM4. Datagrunnlag alternativ 1.

(Konfidensintervall i tabell 3.6)

År for innhenting av data	Funn trinn 1 MaksAvPoeng > 0,3	Kommune Komsentral og Komtjenesteyting	Relativ frekvens	Estimerte avdekking-1-sannsynligheter PM3	 PM4
2007	Funn	Ks=1, Ktj=0	0,50 (2/4)	0,39	0,37
	Funn	Ks=Ktj	0,83 (5/6)	0,61	0,37
	Funn	Ks=0, Ktj=1	1,00 (1/1)	0,80	0,37
	Ikke funn	Ks=1, Ktj=0	0,00 (0/6)	0,16	0,37
	Ikke funn	Ks=Ktj	0,30 (8/27)	0,33	0,37
	Ikke funn	Ks=0, Ktj=1	0,50 (2/4)	0,55	0,37
2006	Funn	Ks=1, Ktj=0	0,00 (0/1)	0,01	0,09
	Funn	Ks=Ktj	0,00 (0/6)	0,21	0,09
	Funn	Ks=0, Ktj=1	0,00 (0/1)	0,41	0,09
	Ikke funn	Ks=1, Ktj=0	0,09 (1/11)	0,03	0,09
	Ikke funn	Ks=Ktj	0,09 (3/33)	0,08	0,09
	Ikke funn	Ks=0, Ktj=1	0,50 (1/2)	0,17	0,09
		Total	0,22 (23/103)		

Vi merker oss at om vi legger prediksjonsmodell 3 til grunn, så synes det å ha vært en økning i avdekking-1-sannsynlighetene fra 2006 til 2007, som gir noe evidens til muligheten at det har skjedd en utvikling/endring fra 2006 til 2007 av praksis når det gjelder gjennomføringen av trinn-1-kontrollene. Det kan naturligvis også være andre årsaker til dette. Til støtte for "andre årsaker" er den logistiske analysen i avsnitt 5 som viser at T ikke inngår blant de signifikante forklaringsvariablene for Z_{mav}. Når det gjelder Z_{sv} derimot, viser analysen i avsnitt 5 at det kan ha vært en utvikling i praksis. Andre årsaker er naturligvis også mulig her.

Kommunetypen der det er størst sjanse for avdekking-1 er i henhold til PM3 ikke-sentrale kommuner som først og fremst er tjenesteytende.

3.2 Sannsynligheten for avdekking-1 (av typen "sum-aarsak"). Datagrunnlag alternativ 2.

Tilsvarende utvikling for alternativ 2 er gitt i tabell 3.4. Merk at siden *Engros* + *Rengj* = 1 eksakt i alternativ 2, forekommer bare *Engros* blant forklaringsvariablene.

Tabell 3.4 Regresjonsresultater (logistisk regresjon)for avdekking av typen "sum-aarsak". Sammenslått materiale. Datagrunnlag alternativ 2.

(Basert på utskrift Ut 6-8 i appendiks A2.1)

Avhengig Y_1	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
T	1.5893	0.013	1.7782	0.002	1.7388	0.002
Zmav	0.5720	0.536	1.2672	0.036	----	----
Zsv	0.5969	0.428	----	----	----	----
Engros	0.1808	0.779	----	----	----	----
Nyreg	0.8505	0.187	----	----	----	----
ENK	0.5188	0.474	----	----	----	----
Ansatte0	0.0781	0.908	----	----	----	----
R	-0.6373	0.309	----	----	----	----
Komsentral	-0.9948	0.128	----	----	----	----
KomTjenest	1.0658	0.099	----	----	----	----
KSminKTJ	----	----	-0.9910	0.059	----	----
konstant	-2.6907	0.001	-2.5193	0.000	-2.2824	0.000
Antall obs	102		102		103	
Log-likelihood	-42.1638		-44.672		-48.8776	
-2 log LR			5.0164		13.4276	
P-verdi redusert vs full modell			0.658		0.144	

Vi ser at bransjen *Engros* falt ut her i motsetning til alternativ 1. Dette er naturlig siden alternativ 1 viser at effekten av *Engros* og *Rengjøring* er nesten like.

På samme måte som under alternativ 1 er både PM1 og PM2 aktuelle. Ingen kan forkastes mot den fulle modellen med en LR-test, og AIC og BIC peker ut hver sin.

Om vi legger til Z_{sv} i PM1, blir ingen av Z_{mav} og Z_{sv} signifikante. Mao., når det gjelder sannsynligheten for avdekking-1, synes ikke Z_{sv} å ha informasjon utover Z_{mav} (se imidlertid avsnitt 3.5).

I tillegg er det interessant å legge merke til at effekten av Z_{mav} (lik *Z* i HRS) riktignok er nær signifikant, men mer diffus enn effekten av den tilsvarende *Z* i HRS der den var temmelig klar. Det er vanskelig å vite hva dette skyldes, men en mulighet kunne være at det har skjedd en utvikling/endring fra 2006 til 2007 av praksis når det gjelder gjennomføringen av trinn-1-kontrollene slik at resultatene fra 2006 snarere bidrar til å tilsløre effekten enn å gjøre den klarere. Heldigvis viser analysen i avsnitt 5 at det er lite evidens i data for denne muligheten når det gjelder Z_{mav}. En annen ting er at *Z*-ene indikerer muligheten for uregelmessigheter av alle typer og er ikke spesielt rettet mot avdekking-1. Dette kan bidra til å tilsløre effekten.

For øvrig er resultatene nokså like som under alternativ 1, og alternativ 1 virker ikke å bidra vesentlig utover alternativ 2 når det gjelder sannsynligheten for avdekking-1.

Tabell 3.5 Estimerte avdekking-1-sannsynligheter, betinget av Z_{mav}, for bransjene *Engros* og *Rengjøring* slått sammen. Basert på prediksjonsmodell PM1 og PM2. Datagrunnlag alternativ 2.

(Konfidensintervall i tabell 3.6)

År for innhenting av data	Funn trinn 1 MaksAvPoeng > 0.3	Kommune Komsentral og Komtjenesteyting	Relativ frekvens	Estimerte avdekking-1-sannsynligheter PM1	 PM2
2007	Funn	Ks=1, Ktj=0	0.50 (2/4)	0.39	0.37
		Ks=Ktj	0.83 (5/6)	0.63	0.37
		Ks=0, Ktj=1	1.00 (1/1)	0.82	0.37
	Ikke funn	Ks=1, Ktj=0	0.00 (0/6)	0.15	0.37
		Ks=Ktj	0.30 (8/27)	0.32	0.37
		Ks=0, Ktj=1	0.50 (2/4)	0.56	0.37
2006	Funn	Ks=1, Ktj=0	0.00 (0/1)	0.01	0.09
		Ks=Ktj	0.00 (0/6)	0.22	0.09
		Ks=0, Ktj=1	0.00 (0/1)	0.44	0.09
	Ikke funn	Ks=1, Ktj=0	0.09 (1/11)	0.03	0.09
		Ks=Ktj	0.09 (3/33)	0.08	0.09
		Ks=0, Ktj=1	0.50 (1/2)	0.18	0.09
		Total	0.22 (23/103)		

Vi ser at det er bare ubetydelige forskjeller (i annen desimal) i estimatene for PM3 (alternativ 1) og den tilsvarende PM1 for alternativ 2.

3.3 Sammenligning av datagrunnlagene alternativ 1 og 2

Problemet er hvilket datagrunnlag (alternativ 1 eller 2) som er mest informativt. Da kan det lønne seg å se på standardfeilen for estimerte størrelser som blant annet er et uttrykk for informasjonsmengden i data. Standardfeilen på estimerte sannsynligheter gir ikke spesielt god mening på grunn av den ikke-lineære skalaen for sannsynligheter. Bedre er å ta utgangspunkt i standardfeilen på de estimerte logit-ene og se hvordan de avspeiler seg i konfidensintervall for sannsynlighetene. Dette er vist i tabell 3.6a og 3.6b som viser konfidensintervallene.

Tabell 3.6a 95% konfidensintervall for avdekking-1-sannsynlighetene, betinget av Z_{mav}, for PM3 under alternativ 1 og for PM1 under alternativ 2 (fra tabell 3.1b og 3.4)

År	Funn trinn 1 Mav > 0.3	Kommune[14]	Alternativ 1			Alternativ 2		
				95% KI			95% KI	
			Sanns.het	Nedre	Øvre	Sanns.het	Nedre	Øvre
2007	Funn	Ks=1, Ktj=0	0.39	0.14	0.70	0.39	0.14	0.71
		Ks=Ktj	0.61	0.35	0.82	0.63	0.36	0.83
		Ks=0, Ktj=1	0.80	0.46	0.95	0.82	0.47	0.96
	Ikke funn	Ks=1, Ktj=0	0.16	0.06	0.38	0.15	0.05	0.38
		Ks=Ktj	0.33	0.20	0.49	0.32	0.19	0.48
		Ks=0, Ktj=1	0.55	0.27	0.80	0.56	0.27	0.82
2006	Funn	Ks=1, Ktj=0	0.01	0.02	0.33	0.01	0.02	0.33
		Ks=Ktj	0.21	0.07	0.48	0.22	0.08	0.50
		Ks=0, Ktj=1	0.41	0.12	0.77	0.44	0.13	0.80
	Ikke funn	Ks=1, Ktj=0	0.03	0.01	0.12	0.03	0.01	0.12
		Ks=Ktj	0.08	0.03	0.19	0.08	0.03	0.18
		Ks=0, Ktj=1	0.17	0.05	0.45	0.18	0.05	0.46

Tabell 3.6b 95% konfidensintervall for avdekking-1-sannsynlighetene, betinget av Z_{mav}, for PM4 under alternativ 1 og for PM2 under alternativ 2 (fra tabell 3.1b og 3.4)

	Alternativ 1			Alternativ 2		
		95% KI			95% KI	
År	Sanns.het	Nedre	Øvre	Sanns.het	Nedre	Øvre
2007	0.37	0.25	0.51	0.37	0.25	0.51
2006	0.09	0.04	0.20	0.09	0.04	0.20

Vi ser at det er svært liten forskjell blant resultatene både for sannsynligheten og deres konfidensintervall mellom alternativ 1 og 2 (i tabell 3.6b ingen forskjell i det hele tatt). Der hvor det er forskjell i konfidensintervallene er intervallene for alternativ 2 gjerne en tanke lengre. Dette kan tolkes slik at noe informasjon blir tapt når man i alternativ 2 kutter ut observasjoner fra andre bransjer enn *Engros* og *Rengjøring*. I utgangspunktet ble observasjoner fra bransjen *Databehandling* fra pilot-dataene innhentet i 2006 kuttet i denne analysen, men resultatet i tabell 3.6 tyder på at det informasjonstapet er neglisjerbart.

3.4 Sannsynligheten for avdekking av typen "*endringer - alle typer*"

Tilsvarende regresjonsresultater for avdekking av typen "*endringer - alle typer*" ($Y = Y_2$) er gitt for alternativ 1 i tabell 3.7-8, basert på utskrift Ut 9-11. Når det gjelder *T* og Z_{mav} er bildet relativt likt som for avdekking-1. Forskjellen er først og fremst at variabelen *Nyregistrert* kommer fram som potensielt betydningsfull for avdekking på trinn 2 i motsetning til avdekking-

[14] Ks står for variabelen *Komsentral* og Ktj for *Komtjenest*

1 samt at *kommunesentralitet* og *tjenesteyting* ikke lenger synes å ha betydning. Det at *Nyreg* slår ut for $Y = Y_2$, men ikke for Y_1, kan ha sammenheng med at "feilaktig bruk av mva-satser" inngår i Y_2 men ikke i Y_1.

På samme måte som for avdekking-1 er det foreslått to aktuelle prediksjonsmodeller. Ingen av dem kan forkastes mot den fulle modellen. Som for avdekking-1, er både Z_{mav} og Z_{sv} svakt signifikante eller nær signifikante om de opptrer enkeltvis, men ikke om de opptrer sammen. For sammenligningens skyld er Z_{mav} prioritert i en eventuell prediksjonsmodell.

Avsnitt 3.5 inneholder en diskusjon av Z_{sv}'s betydning utover Z_{mav} for avdekking-sannsynligheten. Selv om de sammen ikke er signifikante, vil likevel studiet av p-verdiene fra noen beslektete prediksjonsmodeller indikere at Z_{sv} har en viss betydning utover Z_{mav} (for bransjene *Engros* og *Rengjøring*).

Tabell 3.7 Regresjonsresultater (logistisk regresjon) for avdekking av typen "endringer - alle typer". Datagrunnlag alternativ 1.

(Basert på utskrift Ut 9-11 i appendiks A2.2)

Avhengig $Y = Y_2$	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
T	1.2699	0.022	1.3634	0.009	1.5327	0.002
Zmav	0.6875	0.375	1.0439	0.067	----	----
Zsv	0.6423	0.321	----	----	----	----
Engros	1.9093	0.037	----	----	----	----
Rengj	1.7658	0.040	----	----	----	----
EngrosRengj	----	----	1.9800	0.014	2.4010	0.002
Nyreg	0.8674	0.128	0.9470	0.065	----	----
ENK	0.3645	0.563	----	----	----	----
Ansatte0	-0.5152	0.388	----	----	----	----
R	-0.1448	0.800	----	----	----	----
Komsentral	-0.6193	0.250	----	----	----	----
KomTjenest	0.5381	0.336	----	----	----	----
konstant	-3.9213	0.000	-4.2742	0.000	-4.3053	0.000
Antall obs	135		135		137	
Log-likelihood	-55.7294		-57.5128		-61.5662	
-2 log LR			3.5668		11.6736	
P-verdi redusert vs full modell			0.828		0.232	

Tabell 3.8a Sannsynligheter for avdekking av typen "endringer - alle typer" , betinget av Z_{mav}, for prediksjonsmodell 1 (fra tabell 3.7). Datagrunnlag alternativ 1.

År	Funn trinn 1 Mav > 0.3	Nyregistrert	Rel. frekvens	Sanns.het	95% KI Nedre	Øvre
2007	Funn	Ja	1.00 (5/5)	0.74	0.46	0.91
		Nei	0.50 (3/6)	0.53	0.27	0.77
	Ikke funn	Ja	0.42 (5/12)	0.50	0.29	0.72
		Nei	0.28 (7/25)	0.28	0.16	0.46
2006	Funn	Ja	1.00 (1/1)	0.42	0.15	0.75
		Nei	0.00 (0/7)	0.22	0.08	0.48
	Ikke funn	Ja	0.14 (1/7)	0.21	0.08	0.45
		Nei	0.13 (5/39)	0.09	0.04	0.20
		Total	0.26 (27/102)			

Tabell 3.8b Sannsynligheter for avdekking av typen "endringer - alle typer" , betinget av Z_{mav}, for prediksjonsmodell 2 (fra tabell 3.7). Datagrunnlag alternativ 1.

År	Funn trinn 1 Mav > 0.3	Nyregistrert	Rel. frekvens	Sanns.het	95% KI Nedre	Øvre
2007	Funn	Ja	0.42 (20/48)	0.41	0.28	0.55
2006	Ikke funn	Nei	0.13 (7/54)	0.13	0.06	0.25
		Total	0.26 (27/102)			

3.5 Gir kriteriet *samlet vurdering* (811) informasjon utover kriteriet *MaksAvPoeng* > 0.3 for avdekking-sannsynligheten på trinn 2?

I HRS framkom noe evidens for at funn på trinn 1 av typen *samlet vurdering* (811) inneholder informasjon utover funn av typen *MaksAvPoeng* > 0.3 for sannsynligheten for avdekking på trinn 2. Informasjonsmengden i data er tydeligvis ikke sterk nok til at variablene Z_{mav} og Z_{sv} blir signifikante når de opptrer sammen i en regresjonsmodell for avdekkings-sannsynligheten selv om de hver for seg er signifikante når de opptrer alene. Likevel synes tendensen til at Z_{sv} har informasjon utover Z_{mav} rimelig klar hvis vi sammenligner effekten av dem, uttrykt ved p-verdier, i noen beslektete prediksjonsmodeller med forskjellige kombinasjoner av Z_{mav} og Z_{sv} med i modellen, som vist i tabell 3.9. Tendensen synes klarest for alternativ 2 som derfor er valgt som datagrunnlag.

Tabell 3.9 P-verdier fra noen alternative prediksjonsmodeller for *Y* med ulike kombinasjoner av Z_{mav} og Z_{sv}. Datagrunnlag alternativ 2. "----" indikerer at tilsvarende kovariat ikke er med.

(Fra utskrift Ut 12-15 i appendiks A2.2)

Modell	1	2	3	4
	Z_{mav}	Z_{sv}	Z_{mav} og Z_{sv}	Z_{mav} og Z_{altsv}
T	0.010	0.026	0.020	0.017
Zmav	0.061	-----	0.388	0.047
Zsv	-----	0.040	0.265	-----
Zaltsv	-----	-----	-----	0.504
Nyreg	0.049	0.114	0.105	0.079
konstant	0.000	0.000	0.000	0.000

Modell 1 svarer til PM1 i tabell 3.7. Effekten av Z_{mav} og *Nyreg* kommer litt klarere fram under alternativ 2 selv om signifikansen til Z_{mav} fortsatt er noe diffus.

Signifikansen til Z_{sv}, derimot, synes klarere i modell 2. Til gjengjeld "forsvinner" signifikansen til *Nyreg*. Det virker som om kriteriet *samlet vurdering* (811) fokuserer mer på faktoren *Nyregistrert* enn kriteriet *MaksAvPoeng* slik at noe av effekten av *Nyreg* i modell 1 så å si blir overtatt av Z_{sv} i modell 2. Denne tolkningen støttes også av den logistiske analysen av $(Z_{sv} \mid Z_{mav})$ i avsnitt 5.

I modell 3 ser vi at både Z_{mav} og Z_{sv} mister sin signifikans dersom de begge er med. I tillegg forsvinner effekten av *Nyreg*, men det skyldes antakelig at Z_{sv} er med.

I modell 4 har jeg erstattet Z_{sv} med dummien Z_{altsv} som er lik 1 når både $Z_{mav} = 0$ og $Z_{sv} = 1$, og lik 0 ellers, ut fra tanken om at eventuell tilleggsinformasjon fra *samlet vurdering* (811) kanskje først og fremst gjør seg gjeldende når kriteriet *MaksAvPoeng* > 0.3 ikke slår til. Vi ser at effekten av Z_{mav} nå blir klarere enn i modell 1 og at effekten av *Nyreg* delvis blir restaurert, selv om Z_{altsv} selv langt fra er signifikant. Dette siste er trolig et utslag av at datagrunnlaget er litt tynt for dette spørsmålet og antyder samtidig på hvilken måte materialet er tynt, nemlig ved at kombinasjonen $Z_{mav} = 0$ og $Z_{sv} = 1$ kun er representert ved 5 observasjoner (jfr. tabell 4.4).

I tillegg framkommer noe evidens nedenfor i avsnitt 4 at kriteriet *samlet vurdering* (811) kan bidra til å finne fram til "mindre alvorlige endringstilfeller", målt med endringsbeløpets størrelse, enn kriteriet *MaksAvPoeng* > 0.3 i situasjoner der det siste kriteriet ikke slår til.

4. Betydning av andre kovariater og funn på trinn 1 og for forventet endringstall gitt avdekking-1 (sum-aarsak)

4.1 Datagrunnlag alternativ 1

En tilsvarende analyse som i avsnitt 9 i HRS leder til tabell 4.1 som angir full-modell og to alternative prediksjonsmodeller. Dette leder til forslaget om en tredje prediksjonsmodell i tabell 4.3 som den foretrukne under alternativ 1.

Tabell 4.1 GLM-analyse av forventet endringstall gitt avdekking (Alternativ 1)

(Fra utskrift Ut 16-18 i appendiks A2.3)

Avhengig X1	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
(sum-aarsak)	Koeff	P-verdi	Koeff	P-verdi	Koeff	P-verdi
T	-0.3157	0.629	-----	-----	-----	-----
Zsv	-1.2573	0.146	-1.5395	0.012	-----	-----
Zmav	1.6029	0.067	1.0040	0.052	-----	-----
Engros	-2.8225	0.066	-----	-----	-----	-----
Rengj	-2.5870	0.070	-----	-----	-----	-----
Nyreg	0.8691	0.187	0.9353	0.054	-----	-----
ENK	-0.4493	0.533	-----	-----	-----	-----
Ansatte0	-0.5524	0.419	-----	-----	-----	-----
R	-1.0501	0.056	-0.8003	0.066	-----	-----
Komsentral	-0.4504	0.418	-----	-----	-----	-----
KomTjenest	-0.4244	0.506	-----	-----	-----	-----
konstant	16.3418	0.000	12.7612	0.000	12.2892	0.000
Antall obs	24		24		24	
Dispersjon	1.051539		0.8177796		1.330093	
Log-likelihood	-312.543		-314.835		-318.941	
Devians	18.5417		23.1252		31.3366	
AIC	27.04		26.65		26.66	
BIC	-19.59		-37.26		-41.76	
p-verdi redusert vs full modell			0.711		0.307	

Det er ingen av kovariatene i den fulle modellen som klart peker seg ut skal være med. Hvilken av de to prediksjonsmodellene (PM) som er best er heller ikke klart. BIC-kriteriet[15] peker ut PM2 som er uten kovariater, mens AIC svakt favoriserer PM1, men skiller ikke klart mellom de to. Likelihood-ratio (LR) testing av PM2 mot PM1 gir en p-verdi på 0.084, som heller ikke skiller klart mellom dem.

PM1 er interessant siden den indikerer en motsatt tendens fra analysen av avdekking-sannsynlighetene, nemlig at *Z*-ene (Z_{sv} og Z_{mav}) *sammen* synes å ha betydning her. Hver for seg

[15] Aktuelle kandidater basert på informasjonskriteriene AIC og BIC er de som har minst verdier på disse.

synes betydningen å forsvinne, noe tabell 4.2 vitner om ved fordelingen av p-verdier for noen submodeller av PM1.

Tabell 4.2 P-verdi-fordeling for noen submodeller av PM1. "----" indikerer at tilsvarende kovariat ikke er med.

	Kovariater						
Modell	Zmav	Zsv	Nyreg	R	Devians	AIC	BIC
1	0.052	0.012	0.054	0.066	23.12	26.65	-37.26
2	0.727	----	0.485	0.317	27.72	26.76	-35.84
3	----	0.140	0.071	0.089	25.90	26.69	-37.66
4	0.168	----	----	----	29.37	26.66	-40.55
5	----	0.747	----	----	31.20	26.74	-38.71
6	----	----	----	----	31.34	26.66	-41.76

Det er altså ikke noe å hente fra submodeller av PM1 bortsett fra PM2 som impliserer samme forventet endringsbeløp for alle verdier av det vi kalte andre kovariater.

Imidlertid er det et annet problem med PM1 som diskvalifiserer denne modellen og diskuteres nedenfor i tilknytning til tabell 4.4. En mulig løsning på det problemet er å erstatte Z_{sv} med den alternative dummien for definert ved

$$Z_{altsv} = (1 - Z_{mav}) \cdot Z_{sv}$$

Dette er en dummy som kan brukes til å måle innflytelsen av "*samlet vurdering* (811)" utover "*MaksAvPoeng*" når $MAV \leq 0.3$ (dvs. når $Z_{mav} = 0$) som diskutert i avsnitt 3.

Tabell 4.3 viser en tilsvarende analyse som tabell 4.1 der Z_{sv} er erstattet med Z_{altsv}.

Tabell 4.3 GLM-analyse av forventet endringstall gitt avdekking med Z_{mav} og Z_{altsv} istedenfor Z_{sv}. (Alternativ 1)

(Fra utskrift Ut 19-20 i appendiks A2.3)

Avhengig X_1	Full modell		Prediksjonsmodell 3	
	Koeff	P-verdi	Koeff	P-verdi
T	-0.3157	0.629	----	----
Zmav	0.3456	0.708	----	----
Zaltsv	-1.2573	0.146	-1.0396	0.025
Engros	-2.8225	0.066	----	----
Rengj	-2.5870	0.070	----	----
Nyreg	0.8691	0.187	0.7402	0.053
ENK	-0.4493	0.533	----	----
Ansatte0	-0.5524	0.419	----	----
R	-1.0501	0.056	----	----
Komsentral	-0.4504	0.418	----	----
KomTjenest	-0.4244	0.506	----	----
_konstant	16.3418	0.000	12.0868	0.000
Antall obs	24		24	
Dispersjon	1.051539		0.795785	
Log-likelihood	-312.543		-316.2879	
Devians	18.5418		26.0310	
AIC	27.05		26.61	
BIC	-19.59		-40.71	
p-verdi redusert vs full modell			0.586	

Vi er igjen i situasjonen med tre prediksjonsmodeller som alle er aktuelle ut fra vanlige statistiske kriterier. Ingen av dem kan forkastes ut fra LR-tester. AIC og BIC peker ut hver sin av PM1 og PM2. Etter min mening er imidlertid PM3 den mest aktuelle på grunn av følgende diskusjon:

La $\mu_U(Z_{mav}, Z_{sv})$ betegne forventet endringsbeløp i den betingete fordelingen for endringsbeløpet gitt *avdekking*-1, Z_{mav}, Z_{sv} og U [16], der *U* betegner vektoren av andre kovariater som inngår. Tabell 4.4 gir en oversikt over gjennomsnittlige endringstall for forskjellige kategorier av Z_{mav} og Z_{sv}, samt tilsvarende estimerte gjennomsnittlige forventningsverdier under de tre aktuelle prediksjonsmodellene.

[16] Skrevet symbolsk $\mu_U(Z_{mav}, Z_{sv}) = E(X_1 \mid Y_1 = 1, Z_{mav}, Z_{sv}, U)$

Tabell 4.4 Gjennomsnittlig observert endringsbeløp gitt avdekking-1 og gjennomsnittlig estimert forventet endringsbeløp basert på prediksjonmodell PM1-3 fra tabell 4.1 og 4.2.

Z_{mav}	Z_{sv}	Gjennomsnittlig endringsbeløp i data		Gjennomsnittlig estimat av $\mu_U(Z_{mav}, Z_{sv})$ over U		
		Antall obs.	$\bar{X}_1$	PM1	PM2	PM3
1	1	8	309 477	233 446	217 336	220 718
1	0	0	0	1 088 426	217 336	220 718
0	1	5	108 380	85 534	217 336	78 044
0	0	11	170 256	398 796	217 336	220 718

Vi ser at i den andre kategorien der $Z_{mav} = 1$ og $Z_{sv} = 0$, er det ingen avdekking av typen "sum-aarsak" samtidig som PM1 gir et påfallende høyt anslag. Dette virker urimelig siden det ikke er informasjon i data fra denne kategorien. Mer rimelig vil det derfor være å postulere for eksempel at forventet endringsbeløp gitt avdekking når $Z_{mav} = 1$, er det samme enten Z_{sv} er 0 eller 1. Det er nettopp det som oppnås ved å erstatte Z_{sv} med Z_{altsv}.

Det estimerte gjennomsnittsbeløpet i tredje kategori der $Z_{mav} = 0$ og $Z_{sv} = 1$, er lavere enn for de andre kategoriene under PM3. Dette skyldes at koeffisienten for Z_{altsv} under PM3 er negativ. En mulig tolkning av dette er kanskje en tendens til at tilfeller som ikke fanges opp av MAV-kriteriet, men av "samlet vurdering"-kriteriet, er knyttet til mindre alvorlige tilfeller målt ved endringsbeløpets størrelse.

At den fjerde kategorien har samme estimert gjennomsnittsbeløp som de to første skyldes at Z_{mav} falt ut under reduksjonen til prediksjonsmodellen. Hvis Z_{mav} hadde vært signifikant, ville det fjerde beløpet vært forskjellig fra de to første.

Det er litt påfallende at *R* (ekstern regnskapsfører) falt ut i PM3 siden *R* spilte en betydelig rolle for endringsbeløpenes størrelse i HRS (som riktignok kun dreide seg om endringer av "alle typer"). Dette kan skyldes "støy" fra data fra andre bransjer enn *Engros* og *Rengjøring* under alternativ 1. *R* dukker nemlig opp igjen i prediksjonsmodellene under alternativ 2.

4.2 Datagrunnlag alternativ 2

Om vi legger datagrunnlag 2 til grunn framkommer ingen nye effekter, men effektene diskutert i avsnitt 4.1 synes klarere samt at den "tapte" effekten av *R* (ekstern regnskapsfører) dukker opp igjen. Når det gjelder analysen av endringsbeløpenes størrelse synes det økte datagrunnlaget under alternativ 1 således ikke å bidra med annet enn tilført støy.

Tilsvarende tabell 4.1 får vi:

Tabell 4.5 GLM-analyse av forventet endringstall gitt avdekking - med Zmav og Zsv. Alternativ 2.

(Fra utskrift Ut 21-23 i appendiks A2.3)

Avhengig X1 (uten Zaltsv)	Full modell		Prediksjonsmodell 1		Prediksjonsmodell 2	
	Koeff.	p-verdi	Koeff.	p-verdi	Koeff.	p-verdi
T	-0.3157	0.629	-----	-----	-----	-----
Zmav	1.6029	0.067	1.0082	0.039	-----	-----
Zsv	-1.2573	0.146	-1.3810	0.015	-----	-----
Engros	-0.2355	0.691	-----	-----	-----	-----
Nyreg	0.8691	0.187	1.0164	0.023	-----	-----
ENK	-0.4493	0.533	-----	-----	-----	-----
Ansatte0	-0.5524	0.419	-----	-----	-----	-----
R	-1.0501	0.056	-0.8905	0.028	-----	-----
Komsentral	-0.4504	0.418	-----	-----	-----	-----
KomTjenest	-0.4244	0.506	-----	-----	-----	-----
konstant	13.7548	0.000	12.5889	0.000	12.2334	0.000
Antall obs	23		23		23	
Dispersjon	1.051539		0.733166		1.472102	
Log-likelihood	-298.4436		-299.46		-304.3674	
Devians	18.5418		20.5778		30.38919648	
AIC	26.91		26.47		26.55	
BIC	-19.08		-35.86		-38.59	
p-verdi redusert vs full modell			0.916		0.295	

Tilsvarende tabell 4.3 får vi:

Tabell 4.5 GLM-analyse av forventet endringstall gitt avdekking - med Zmav og Zaltsv istedenfor Zsv. Alternativ 2.

(Fra utskrift Ut 24-25 i appendiks A2.3)

Avhengig X1 (med Zaltsv)	Full modell		Prediksjonsmodell 3	
	Koeff.	p-verdi	Koeff.	p-verdi
T	-0.3157	0.629	-----	-----
Zmav	0.3456	0.708	-----	-----
Zaltsv	-1.2573	0.146	-1.1404	0.013
Engros	-0.2355	0.691	-----	-----
Nyreg	0.8691	0.187	0.8355	0.027
ENK	-0.4493	0.533	-----	-----
Ansatte0	-0.5524	0.419	-----	-----
R	-1.0501	0.056	-0.7901	0.038
Komsentral	-0.4504	0.418	-----	-----
KomTjenest	-0.4244	0.506	-----	-----
konstant	13.7548	0.000	12.4284	0.000
Antall obs	23		23	
Dispersjon	1.051539		0.72709	
Log-likelihood	-298.4437		-299.6551	
Devians	18.5418		20.9647	
AIC	26.91		26.40	
BIC	-19.08		-38.61	
p-verdi redusert vs full modell			0.933	

Som før kan ingen av de tre prediksjonsmodellene (PM1-3) forkastes ved LR-testing. PM1 har samme problem som beskrevet under alternativ 1 og er derfor ikke å anbefale. Kriteriene AIC og BIC peker begge ut PM3 som best. Både *Nyreg* og *R* har klare effekter i PM3.

Den samlete effekten av Z_{mav} og Z_{sv} er i PM3 redusert til effekten av Z_{altsv} alene uten vesentlig tap i tilpasningsgrad. Den estimerte effekten av Z_{altsv} er negativ som impliserer mindre forventet endringsbeløp i situasjoner der kriteriet *MaksAvPoeng* > 0.3 ikke slår til samtidig som kriteriet basert på revisors *samlet vurdering* (811) slår til med høy verdi (3 eller 4).

Således er PM3 et klart valg i dette tilfellet. PM2 kan også eventuelt brukes til å beregne gjennomsnittsbeløp som PM2 i tabell 4.4.

5. Sannsynlighetsfordlingen for funn på trinn 1

Det er to responsvariable for funn på trinn 1, Z_{mav} og Z_{sv}. De inngår i de betingete fordelingene for Y_1 og X_1, så vi trenger å modellere den simultane fordelingen for Z_{mav} og Z_{sv} (de er åpenbart ikke uavhengige) for å få en full sannsynlighetsmodell for responsen, $(Z_{mav}, Z_{sv}, Y_1, X_1)$. Dette krever en viss modifikasjon av formelverket utviklet i HRS siden responsen der var 3-dimensjonal, (Z, Y, X). Modifikasjonen er utført i appendiks A1.

For den simultane fordelingen til Z_{mav} og Z_{sv} har vi to naturlige valg i tråd med den generelle metodikken brukt her: (i) multippel logit eller (ii) to univariate logit-analyser basert på multiplikasjonssetningen for sannsynligheter:

$$P(Z_{mav} = z_1, Z_{sv} = z_2 \mid U) = P(Z_{mav} = z_1 \mid U) \cdot P(Z_{sv} = z_2 \mid Z_{mav} = z_1, U)$$

der *U* er vektoren av andre (eksogene) kovariater. Den andre veien der Z_{mav} og Z_{sv} bytter plass på høyre side, er også mulig og gir en tredje variant, men den foreslåtte veien er valgt siden den har større substansiell interesse.

(i) og (ii) gir to forskjellige, ikke helt ekvivalente, men ganske like simultane fordelinger. Siden STATA har implementert multippel logit, prøvde jeg begge, men fant at (ii) ga best tilpasning i tillegg til at (ii) er enklere å bruke.

Dette betyr to vanlige univariate logit-analyser for $Z_{sv} \mid Z_{mav}$ og Z_{mav} henholdsvis, referert i tabell 5.1 og 5.2.

Bare resultatene basert på datagrunnlaget alternativ 2 er rapportert her siden alternativ 1 ga samme redusert modell med små forskjeller, og alternativ 2, først og fremst på grunn av resultatene i avsnitt 4, er valgt som grunnlag for anvendelser av modellen i avsnitt 6.

Tabell 5.1 Regresjonsresultater (logistisk regresjon) for funn på trinn 1 av typen "samlet vurdering", gitt Z_{mav}. Datagrunnlag alternativ 2.

(Basert på utskrift Ut 26-27 i appendiks A2.4)

Avhengig $Z_{sv} \mid Z_{mav}$	Full modell		Prediksjonsmodell	
	Koeff.	p-verdi	Koeff.	p-verdi
T	1.1336	0.015	1.2115	0.007
Zmav	4.1489	0.000	4.3600	0.000
Engros	0.2647	0.545	-----	-----
Nyreg	1.1283	0.006	1.1466	0.003
ENK	0.3702	0.446	-----	-----
Ansatte0	0.3126	0.475	-----	-----
R	-0.0394	0.932	-----	-----
Komsentral	0.0027	0.995	-----	-----
KomTjenest	0.0539	0.900	-----	-----
konstant	-3.6112	0.000	-3.2608	0.000
Antall obs	293		293	
Log-likelihood	-94.6056		-95.7049	
-2 x log LR			2.1985	
P-verdi			0.901	

Tabell 5.2 Regresjonsresultater (logistisk regresjon) for funn på trinn 1 av typen "MaksAvPoeng". Datagrunnlag alternativ 2.

(Basert på utskrift Ut 28-29 i appendiks A2.4)

Z_{mav} (uten Z_{sv})	Full modell		Prediksjonsmodell	
	Koeff.	p-verdi	Koeff.	p-verdi
T	-0.0764	0.877	-----	-----
Engros	0.1429	0.793	-----	-----
Nyreg	0.4289	0.410	-----	-----
ENK	2.5238	0.000	2.3961	0.000
Ansatte0	0.1661	0.754	-----	-----
R	-0.5797	0.266	-----	-----
Komsentral	0.1426	0.770	-----	-----
KomTjenest	0.1410	0.784	-----	-----
konstant	-4.0343	0.000	-4.0254	0.000
Antall obs	293		293	
Log-likelihood	-68.3732		-69.5314	
-2 x log LR			2.3163	
P-verdi			0.940	

Tabell 5.1 viser klart at Z_{mav} og Z_{sv} er sterkt avhengige samt at Z_{sv} har potensiell tilleggsinformasjon utover Z_{mav} knyttet til faktoren *Nyregistrert* i betydningen at fordelingen for $Z_{sv} \mid Z_{mav}$ endrer seg med *Nyreg*. Diskusjonen i avsnitt 3.5 og 4.2 viser at denne tilleggsinformasjonen både gjelder sannsynligheten for avdekking på trinn 2 og endringsbeløpets (X_1) størrelse.

Det er interessant at årsdummien *T* falt ut fra fordelingen til Z_{mav}. Dette er naturlig å tolke i retning av at samme prosedyre i forbindelse med kriteriet *MaksAvPoeng* har vært fulgt i begge år. Hvis det har skjedd en endring av praksis for funn på trinn 1 fra 2006 til 2007, er den i så fall knyttet til kriteriet *samlet vurdering* (811).

Dette blir å oppfatte som et forbehold i den foreliggende analysen som bygger på antakelsen at det ikke har skjedd noen endring av praksis på trinn 1.

6. Anvendelser av estimert modell

6.1 Beskrivelse av fordelingen til responsvektoren

Responsvektoren er gitt ved $W = (Z_{mav}, Z_{sv}, Y_1, X_1)$ og de eksogene kovariatene gitt ved en vektor *U*. Analysen i avsnitt 3-5 har redusert *U* til

$$U = (T, ENK, Nyreg, R, KSminKTJ)$$

Som i HRS har *W* en komplisert betinget fordeling gitt *U*, $f(w \mid U = u)$, en blanding av kontinuerlige og diskrete elementer beskrevet nærmere i appendiks A1. Analysen i avsnitt 3-5 har resultert i en forenkling av strukturen av *f* beskrevet nærmere i tabell 6.1, som viser hvordan de eksogene kovariatene i *U* influerer på forskjellige deler av fordelingen.

Generelt gjelder at en flerdimensjonal sannsynlighetstetthet (også for blandete fordelinger) kan faktoriseres ved multiplikasjonssetningen for sannsynligheter til et produkt av endimensjonale tettheter:

$$f(w_1, w_2, w_3, w_4 \mid U) = f_1(w_1 \mid U)\, f_2(w_2 \mid w_1, U)\, f_3(w_3 \mid w_1, w_2, U)\, f_4(w_4 \mid w_1, w_2, w_3, U)$$

der $w_1, \ldots, w_4$ svarer til de fire responsvariablene. Når w-ene står på høyre side av " | ", kan de betraktes som andre kovariater.

Tabell 6.1 viser hvordan kovariatene dukker opp i de fire faktorene basert på analysen i avsnitt 3-5.

Tabell 6.1 Oversikt over hvordan fordelingen til responsvektoren avhenger av de andre kovariatene, *U*. Fortegnet indikerer om effekten til kovariaten er positiv eller negativ.

Faktorer i sannsynlighetsford. for responsvariablene	Forklaringsvariable							
				Andre (eksogene) kovariater U				
	Z_{mav}	Z_{sv}	Y_1	T	*ENK*	*Nyreg*	R	*KSminKTJ*
$f(Z_{mav} \mid U)$	----	----	----	---	*+ENK*	----	----	----
$f(Z_{sv} \mid Z_{mav}, U)$	$+ Z_{mav}$	----	----	*+T*	----	*+Nyreg*	----	----
$f(Y_1 \mid Z_{mav}, Z_{sv}, U)$	$+ Z_{mav}$	----	----	*+T*	----	----	----	*– KSminKTJ*
$f(X_1 \mid Z_{mav}, Z_{sv}, Y_1 = 1, U)$	$-(1 - Z_{mav}) \cdot Z_{sv}$		$Y_1 = 1$	---	----	*+Nyreg*	*– R*	----
$P(X_1 = 0 \mid Z_{mav}, Z_{sv}, Y_1 = 0, U) = 1$ for alle Z og U				---	---	---	---	---

Modellen postulerer videre at de fire faktorene avhenger av fire forskjellige og variasjonsuavhengige sett av parametre som impliserer at full maximum likelihood (ML) analyse av modellen er ekvivalent med ML-analyse av faktorene hver for seg (jfr appendiks A1.2 i HRS), og som er gjennomført i avsnitt 3-5.

6.2 Estimering av forventet endringsbeløp gitt avdekking og sannsynligheten for avdekking

Basert på modellen beskrevet i avsnitt 6.1 har jeg estimert forventet endringsbeløp av typen "sum-aarsak" gitt avdekking-1, symbolsk $E(X_1 \mid U, Y_1 = 1)$, og tilhørende sannsynligheter for avdekking-1, $P(Y_1 = 1 \mid U)$. Effekten av funn på trinn 1 (dvs. av *Z*-ene) er kontrollert for ved at de er aggregert ut av forventning og sannsynlighet som forklart i appendiks A1.

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 har derfor mindre tolkningsmessig interesse. Den er først og fremst nyttig ved aggregering til totaltall over grupper av strata, som gjort rede for i HRS. Den type aggregering er ikke foretatt her, men et uttrykk for $E(X_1 \mid U)$ er utledet i appendiks A1.

Tabell 6.2 Estimater (2006) for forventet endringsbeløp av typen "sum-aarsak" gitt avdekking-1 og sannsynligheten for avdekking-1, kontrollert for funn på trinn 1. Bransje *En gros* og *Rengjøring*. Datagrunnlag alternativ 2.

ENK	Ny-registrert	Ekstern regnskaps-fører	Forventet endringsbeløp gitt avdekking-1 (1000kr.)	Sannsynlighet for avdekking-1 Kommunesentr.(Ks) og Kommunetjenst.(Ktj)		
				Ks=1 Ktj=0	Ks = Ktj	Ks=0 Ktj=1
Ja	Ja	Ja	219	0.189	0.373	0.604
		Nei	482	0.189	0.373	0.604
	Nei	Ja	106	0.189	0.373	0.604
		Nei	234	0.189	0.373	0.604
Nei	Ja	Ja	211	0.154	0.328	0.567
		Nei	465	0.154	0.328	0.567
	Nei	Ja	105	0.154	0.328	0.567
		Nei	231	0.154	0.328	0.567

For sammenligningens skyld, har jeg satt opp en deskriptiv tabell 6.3 over tilsvarende størrelser i data.

Tabell 6.3 Deskriptive resultater for 2006 for gjennomsnittlig endringsbeløp av typen "sum-aarsak" gitt avdekking-1, og relative frekvenser av avdekking-1. Frekvenser i parentes. Bransje *En gros* og *Rengjøring*. Datagrunnlag alternativ 2.

ENK	Ny-registrert	Ekstern regnskaps-fører	Gjennomsnittlige endringsbeløp gitt avdekking-1 i data (1000kr.)	Relative frekvenser for avdekking-1 Kommunesentr.(Ks) og Kommunetjenst.(Ktj)		
				Ks=1 Ktj=0	Ks = Ktj	Ks=0 Ktj=1
Ja	Ja	Ja	162 (5)	1.00 (1/1)	0.67 (4/6)	0 obs.
		Nei	750 (2)	0.00 (0/1)	1.00 (1/1)	1.00 (1/1)
	Nei	Ja	100 (3)	0.00 (0/1)	0.22 (2/7)	0.50 (1/2)
		Nei	114 (2)	0.50 (1/2)	1.00 (1/1)	0 obs.
Nei	Ja	Ja	207 (1)	0.00 (0/1)	0.00 (0/4)	1.00 (1/1)
		Nei	245 (1)	0 obs.	1.00 (1/1)	0 obs.
	Nei	Ja	97 (3)	0.00 (0/2)	0.33 (3/9)	0 obs.
		Nei	314 (1)	0.00 (0/1)	0.33 (1/3)	0.00 (0/1)

Vi ser at avviket mellom estimerte forventninger og sannsynligheter, og tilsvarende gjennomsnitt og relative frekvenser ikke er avskrekkende, men også at datagrunnlaget, spesielt for de relative frekvensene, er for tynt til å se noen klare tendenser.

Tabell 6.2 viser en relativt påfallende høy verdi av forventet endringstall for nyregistrerte virksomheter uten ekstern regnskapsfører. Dette bør tas med en klype salt. I tabell 6.3 finner vi

en uforholdsmessig høy verdi , 750 tusen, for nyregistrerte enkeltmannsvirksomheter uten ekstern regnskapsfører - basert på to observasjoner. Ser vi etter finner vi at en av dem er på 1,2 millioner, som er så ekstrem i forhold til de andre observasjonene av endringstall at den kvalifiserer til betegnelsen "outlier" (eller "uteligger" på norsk) i henhold til vanlig statistisk språkbruk.

Problemet med uteliggere, er at de særlig i små datasett (som vi har her), kan ha utilbørlig stor innflytelse på estimering. Som illustrasjon på fenomenet er gjennomsnittet av de 23 endringstallene, med og uten den ekstreme observasjonen, lik 206 tusen og 160 tusen henholdsvis, som er en betydelig endring forårsaket av en enkelt observasjon.

Således er det all grunn til å tro at de høye verdiene for nyregistrerte virksomheter uten ekstern regnskapsfører, for en stor del skyldes denne ene ekstreme observasjonen. Bekreftelse og utdyping av dette kan finnes i avsnitt 6.3 der jeg har utført to alternative beregninger, den ene ved å fjerne uteliggeren helt, og den andre ved å fjerne en null fra beløpet i tilfelle det skulle være en punchefeil.

Et annet iøynefallende trekk ved tabell 6.2 er de høye sannsynlighetene for avdekking-1, særlig for ikke-sentrale kommuner som hovedsakelig er tjenesteytende. Hvis modellen hadde vært helt sann, ville disse sannsynlighetene kunne oppfattes som prevalens-sannsynligheter. I så fall ville det bety at ca 60% av virksomhetene i slike kommuner driver med skatteunndragelser - som jeg antar virker noe oppsiktsvekkende.

På den annen side er det en god del usikkerhet knyttet til disse sannsynlighetene - ikke minst for den aktuelle kommunetypen som er tynt representert i data (jfr. tabell 6.3). For å nyansere bildet har jeg derfor i tabell 6.4 beregnet ensidige nedre 95% konfidensgrenser[17] for estimatene fra tabell 6.2. Med andre ord, istedenfor å si at sannsynligheten for at en vilkårlig virksomhet unndrar skatt er 0.60 for bransjene *En gros* eller *Rengjøring* i ikke-sentrale kommuner som hovedsakelig er tjenesteytende, er det antakelig bedre å si at vi ikke vet mer om den sannsynligheten enn at den er større enn 0.37 (med konfidens 95% ut fra data og modell).

Standardfeilene som ligger bak konfidensgrensene er beregnet på grunnlag av formler utledet i appendiks A1.2. Standardfeilene er ikke oppgitt her, men kan lett beregnes fra konfidensgrensen ved formelen

$$St.feil = (estimat - konf.grense)/1.645$$

[17] Hvis $\hat{\theta}$ er en estimator for parameteren θ, er en ensidig nedre 95% konfidensgrense en stokastisk variabel, *L*, som oppfyller $P(L \le \theta) = 0.95$. Hvis $\hat{\theta}$ er tilnærmet normalfordelt, har *L* vanligvis formen $L = \hat{\theta} - 1.645 \cdot SE(\hat{\theta})$, der $SE(\hat{\theta})$ er standardfeilen.

Tabell 6.4 Ensidige nedre 95% konfidensgrenser for forventet endringsbeløp av typen "sum-aarsak" gitt avdekking-1 og sannsynligheten for avdekking-1, kontrollert for funn på trinn 1. Bransje *En gros* og *Rengjøring*. Datagrunnlag alternativ 2. Estimatene gjelder 2006.

			Ensidige nedre 95% konfidensgrenser			
			Forventet endringsbeløp gitt avdekking-1	Sannsynlighet for avdekking-1		
				Kommunesentr.(Ks) og Kommunetjenst.(Ktj)		
ENK	Ny-registrert	Ekstern regnskaps-fører	(1000kr.)	Ks=1 Ktj=0	Ks = Ktj	Ks=0 Ktj=1
Ja	Ja	Ja	102	0.048	0.256	0.366
		Nei	171	0.048	0.256	0.366
	Nei	Ja	60	0.048	0.256	0.366
		Nei	96	0.048	0.256	0.366
Nei	Ja	Ja	99	0.023	0.205	0.312
		Nei	168	0.023	0.205	0.312
	Nei	Ja	59	0.023	0.205	0.312
		Nei	95	0.023	0.205	0.312

I lys av den betydelige usikkerheten knyttet til uteliggeren når det gjelder prediksjonsmodell og forventete verdier av X_1, har jeg også beregnet tilsvarende resultater for en forenklet modell (som PM2 i tabell 4.1) der jeg har fjernet alle kovariater fra den betingete fordelingen for X_1. Denne spesifikasjonen kan ikke forkastes med en LR-test relativt til den fulle modellen i tabell 4.3.

Sannsynlighetsestimatene er ikke påvirket av uteliggeren, men usikkerheten synes betydelig i forbindelse med de kommunetypene som har motsatt verdi på *Komsentral* og *Komtjenest*. Jeg har derfor også fjernet avhengigheten av kommunetype i beregningen. Med andre ord, variabelen *KSminKTJ* tas ut av modellen for Y_1 (jfr. tabell 3.4). Heller ikke denne forenklingen kan forkastes mot full-modellen.

Den forenklete modellen er kort oppsummert i tabell 6.5.

Tabell 6.5 Forenklet modell

(Utskrifter i appendiks A2.5)

Avhengig	Kovariater	P-verdi mot Full modell
Y_1	T, Z_{mav}	0.350
X_1	*konstant*	0.295
Z_{mav}, Z_{sv}	*T*, ENK, *Nyreg*	Som før (tabell 5.1-2)

I tabellen som svarer til tabell 6.2, faller *Nyreg* og *R* bort, bare ENK blir igjen. *Nyreg* faller bort selv om den påvirker *Z*-ene, men i den forenklete modellen påvirker ikke *Z*-ene X_1.

Resultatene er samlet i tabell 6.6.

Tabell 6.6 Resultater basert på forenklet modell

ENK	Forventet endringsbeløp gitt avdekking-1 (1000kr.)	Standardfeil (1000kr.)	Ensidig nedre 95% konf. grense	Sanns.het avdekking-1	Standardfeil	Ensidig nedre 95% konf. grense
Ja	206	52	120	0.357	0.069	0.245
Nei	206	52	120	0.318	0.072	0.199

6.3 Effekten av en uteligger

Uteliggeren er en nyregistrert enkeltmannsvirksomhet uten ekstern regnskapsfører og uten ansatte. Den tilhører bransjen *Rengjøring* og ligger i en hovedsakelig tjenesteytende og ikke-sentral kommune. Endringsbeløpet er 1 200 000 kr., mens de øvrige endringsbeløpene ikke overskrider 460 000 kr.

For å se på effekten av denne observasjonen på estimeringene har jeg foretatt to alternative og hypotetiske estimeringer tilsvarende tabell 6.2:

1. En beregning der uteliggeren er fjernet fra data.
2. En beregning der uteliggeren beholdes men beløpet endret til kr. 120 000.

Resultatene er vist i tabell 6.7 og 6.8.

Tabell 6.7 Alternativ beregning 1. Uteligger (verdi 1 200 000 kr.) fjernet. Samme modell som i tabell 6.2.

Estimater (2006) for forventet endringsbeløp av typen "sum-aarsak" gitt avdekking-1 og sannsynligheten for avdekking-1, kontrollert for funn på trinn 1. Bransje *En gros* og *Rengjøring*. Datagrunnlag alternativ 2.

ENK	Ny-registrert	Ekstern regnskaps-fører	Forventet endringsbeløp gitt avdekking-1 (1000kr.)	Sannsynlighet for avdekking-1 Kommunesentr.(Ks) og Kommunetjenst.(Ktj)		
				Ks=1 Ktj=0	Ks = Ktj	Ks=0 Ktj=1
Ja	Ja	Ja	189	0.193	0.367	0.584
		Nei	321	0.193	0.367	0.584
	Nei	Ja	114	0.193	0.367	0.584
		Nei	193	0.193	0.367	0.584
Nei	Ja	Ja	183	0.162	0.326	0.549
		Nei	313	0.162	0.326	0.549
	Nei	Ja	112	0.162	0.326	0.549
		Nei	192	0.162	0.326	0.549

Tabell 6.8 Alternativ beregning 2. Uteligger (verdi kr. 1 200 000) erstattet med kr. 120 000. Samme modell som i tabell 6.2.

Estimater (2006) for forventet endringsbeløp av typen "sum-aarsak" gitt avdekking-1 og sannsynligheten for avdekking-1, kontrollert for funn på trinn 1. Bransje *En gros* og *Rengjøring*. Datagrunnlag alternativ 2.

ENK	Ny-registrert	Ekstern regnskaps-fører	Forventet endringsbeløp gitt avdekking-1 (1000kr.)	Sannsynlighet for avdekking-1 Kommunesentr.(Ks) og Kommunetjenst.(Ktj)		
				Ks=1 Ktj=0	Ks = Ktj	Ks=0 Ktj=1
Ja	Ja	Ja	177	0.189	0.373	0.604
		Nei	266	0.189	0.373	0.604
	Nei	Ja	118	0.189	0.373	0.604
		Nei	178	0.189	0.373	0.604
Nei	Ja	Ja	172	0.154	0.328	0.567
		Nei	259	0.154	0.328	0.567
	Nei	Ja	117	0.154	0.328	0.567
		Nei	176	0.154	0.328	0.567

Vi ser at sannsynlighetene ikke er berørt av uteliggeren. Estimeringen av sannsynlighetene i denne situasjonen er helt uavhengig av størrelsen på X_1-ene. Den lille endringen i tabell 6.7 skyldes bare at vi der har en observasjon mindre. Dette skyldes ganske enkelt at vi kan endre verdiene for X_1 så mye vi vil, så lenge de holder seg over null, uten at datagrunnlaget for Y_1 og *Z*-ene berøres.

Til gjengjeld er det betydelige endringer i de forventete endringsbeløpene. For å lette oversikten har jeg samlet de tre variantene i tabell 6.9.

Tabell 6.9 Forventet endringsbeløp for tre alternative beregninger.

ENK	Ny-registrert	Ekstern regnskaps-fører	Forventet endringsbeløp gitt avdekking-1 (1000kr.) Opprinnelige data	Uteligger fjernet	Uteligger erstattet
Ja	Ja	Ja	219	189	177
		Nei	482	321	266
	Nei	Ja	106	114	118
		Nei	234	193	178
Nei	Ja	Ja	211	183	172
		Nei	465	313	259
	Nei	Ja	105	112	117
		Nei	231	192	176

Resultatene viser at den ene uteliggeren har hatt stor innflytelse på estimatene for forventet endringsbeløp. Resultatene for de hypotetiske beregningene er en god del mer utjevnet. Ikke bare har uteliggeren hevet beløpene betraktelig for nyregistrerte virksomheter uten ekstern regnskapsfører, men også hevet beløpene litt for de andre kategoriene bortsett fra for eldre virksomheter med ekstern regnskapsfører.

Innflytelsen til uteliggeren er dessverre ikke begrenset til beløpsstørrelsene. En databestemt utvikling av prediksjonsmodeller kan også påvirkes av en uteligger. Dette gjelder ikke sannsynlighetene for Y_1 og *Z*-ene som ikke påvirkes av verdien av X_1, men er i høy grad aktuell for fordelingen til X_1. Som illustrasjon har jeg i tabell 6.10 referert resultatet av å utvikle prediksjonsmodeller ut fra data for de to hypotetiske data-settene sammen med den som er basert på de opprinnelige dataene og referert ovenfor.

Tabell 6.10 Prediksjonsmodeller for forventet endringsbeløp gitt avdekking-1, for tre alternative beregninger.

Avhengig X_1	Opprinnelige data		Uteligger fjernet		Uteligger erstattet	
	Koef.	P-verdi	Koef.	P-verdi	Koef.	P-verdi
Zaltsv	-1.1404	0.013	-0.8285	0.074	-0.8366	0.059
Nyreg	0.8355	0.027	-----	-----	-----	-----
Ansatte0	-----	-----	-0.8917	0.051	-0.8751	0.037
R	-0.7901	0.038	-0.6581	0.107	-0.6696	0.082
KomTjenest	-----	-----	-0.9448	0.040	-0.9278	0.028
konstant	12.4284	0.000	13.4229	0.000	13.4196	0.000

Vi ser at modellen endrer seg en del om uteliggeren fjernes eller tones ned. Det viser seg (ikke rapportert) at om vi estimerer den opprinnelige modellen i de hypotetiske situasjonene, forsvinner signifikansen til alle effektene utenom konstantleddet (den som forsvinner minst er effekten til Z_{altsv} som får en p-verdi på 0.09). Vi ser også at, for første gang i denne studien, dukker *Ansatte*0 opp som en betydningsfull kovariat samtidig som *Nyreg* forsvinner, og effekten av ekstern regnskapsfører (*R*) reduseres.

Et naturlig spørsmål som nå melder seg er: Skal den ekstreme observasjonen være med, eller skal den fjernes (eventuelt rettes)? Etter min mening skal den være med - så fremt den er en legitim observasjon. For hvis den er legitim, inneholder den viktig informasjon om populasjonen som studeres, selv om slike observasjoner er sjeldne. Det hører med i rapporten å rapportere om slike og deres eventuelle innflytelse på resultatene. Etter hvert som man får inn mer informasjon fra senere tilfeldige utvalg, vil man kunne si noe om hvor hyppige slike tilfeller er og eventuelt innarbeide denne informasjonen i senere modeller som blant annet kan utnyttes til å kontrollere for ekstreme observasjoner ved senere enkeltundersøkelser. Det er klart mange måter å gjøre dette på. En mulighet kunne, for eksempel være å splitte opp responskategorien "avdekking" i to nye responskategorier, "vanlige" avdekking med endringsbeløp innenfor gitte grenser, og "ekstreme" med endringsbeløp utenfor disse. De nye kategoriene kunne så behandles separat med eventuelt forskjellig statistisk metodikk.

6.4 Effekten av års-indikatoren *T*.

Det har klart begrenset interesse å bruke modellen foreslått her til å estimere forventete endringsbeløp og avdekking-sannsynligheter for 2005 som pilotdataene er hentet fra. For det første er denne undersøkelsen rettet mot 2006, og for det andre er informasjon fra piloten kun hentet i håp om å øke datagrunnlaget for endringsbeløp av typen "sum-aarsak" for 2006.

Likevel kan det ha indirekte interesse å beregne en tilsvarende tabell som tabell 6.2 for 2005, basert på den foreslåtte prediksjonsmodellen for 2006 for å få ut noe informasjon om hvordan års-dummien, *T*, har fungert. En slik diskusjon har også relevans for det første forbeholdet formulert i innledningen.
Hvis nemlig *T* har fungert tilfredsstillende slik at de undertrykte samspillene mellom *T* og andre koveriater (observerte som ikke-observerte) ikke har hatt vesentlig betydning, bør ikke

resultatene for 2005 (vist i tabell 6.11) avvike for mye fra resultatene for 2006. Siden modellen som er brukt i hovedsak er påvirket av dataene fra 2006 og derfor indirekte av forholdene slik de var i 2006, vil store forskjeller kunne tas som evidens for at *T* uten samspill ikke har klart å kontrollere for alle relevante endringer fra 2005 til 2006.

Tabell 6.11 Estimater (2005) for forventet endringsbeløp av typen "sum-aarsak" gitt avdekking-1 og sannsynligheten for avdekking-1, kontrollert for funn på trinn 1. Bransje *En gros* og *Rengjøring*. Datagrunnlag alternativ 2.

ENK	Ny-registrert	Ekstern regnskaps-fører	Forventet endringsbeløp gitt avdekking (1000kr.)	Sannsynlighet for avdekking Kommunesentr.(Ks) og Kommunetjenst.(Ktj)		
				Ks=1 Ktj=0	Ks = Ktj	Ks=0 Ktj=1
Ja	Ja	Ja	245	0.040	0.099	0.220
		Nei	541	0.040	0.099	0.220
	Nei	Ja	111	0.040	0.099	0.220
		Nei	245	0.040	0.099	0.220
Nei	Ja	Ja	243	0.030	0.077	0.183
		Nei	535	0.030	0.077	0.183
	Nei	Ja	111	0.030	0.077	0.183
		Nei	244	0.030	0.077	0.183

To slående trekk kan observeres.

- Beløpsstørrelsene er av samme størrelsesorden som for 2006 som er et godt tegn.
- Avdekking-sannsynlighetene derimot er svært lave i forhold til dem som er estimert for 2006.

Det siste kan naturligvis skyldes en generell svakhet ved modellen slik at sannsynlighetsestimatene er like gale for 2006 som for 2005, men i motsatt retning. Men det er også mulig at forklaringen nettopp skyldes at en enkelt dummy ikke har vært tilstrekkelig for å fange opp alle relevante konjunkturmessige og strukturelle endringer fra 2005 til 2006.

Evidensen her kan altså ikke tas som noe bevis for at *T* ikke har fungert helt ut, snarere en indikasjon for muligheten av det.

7 Noen konklusjoner

- Studien dreier seg hovedsakelig om å estimere for 2006 sannsynligheten for avdekking (funn av endringstilfeller) og forventet endringsbeløp gitt avdekking i et tilfeldig utvalg av virksomheter fra bransjene *En gros* (51 -Engroshandel med klær, sports- og fritidsutstyr mv.) og *Rengjøring* (74). Det ble ikke funnet evidens for vesentlige forskjeller mellom de to bransjene slik at de ble slått sammen i en felles bransjekategori. Endringstilfellene som diskuteres her er av typen "nettoinntekt bortsett fra feilperiodiseringer og feil bruk av mva-satser" i motsetning til hovedrapporten som inkluderte "feil bruk av mva-satser".

- Studien omfatter fire responsvariable, to variable for funn på trinn 1, en dikotom variable for avdekking på trinn 2, samt endringsbeløpets størrelse gitt avdekking på trinn 2.
- Fokus i denne rapporten, som er et supplement til hovedrapporten med et litt annet fokus, er å finne betydningsfulle faktorer ved virksomhetene innenfor en gruppe av bransjer (her *Engros* og *Rengjøring*) og uavhengig av region. De potensielt betydningsfulle faktorene er karakterisert ved 9 dikotome variable som definert i avsnitt 2.3 samt to avledete variable av disse.
- 2006-dataene er trukket stratifisert fra 11 av landets fylker (jfr. tabell 1.2), og supplementert med noen data fra 2005 fra pilotundersøkelsen for å øke informasjonsgrunnlaget. En dummy (*T*) for ble lagt til kovariatene for å kontrollere for strukturelle og konjunkturmessige endringer fra 2005 til 2006. Evidensen tyder på at *T* har fungert rimelig bra når det gjelder endringsbeløpenes størrelse, men store avvik i avdekkingsannsynlighetene mellom 2005 og 2006 tyder på at *T* ikke har klart å fange opp alle relevante endringer.
- To forskjellige datagrunnlag ble prøvet: Alternativ 1 omfatter alle data fra 2006 pluss data fra for bransjene *En gros* og *Rengjøring* fra 2005. Alternativ 2 omfatter alle dataene fra *En gros* og *Rengjøring* for begge årene, men ingen observasjoner fra de andre ikke-overlappende bransjene. Alternativ 1 viste seg ubetydelig mer informativ med hensyn på estimering av avdekking-sannsynligheter, men mer informativ med hensyn på estimering av forventet endringsbeløp. Alternativ 2 ble derfor valgt som grunnlag for anvendelsene i avsnitt 6.
- Alle de foreslåtte kovariatene, med unntak av *Antall ansatte*, viste seg å ha betydning for fordelingen for responsvektoren. At *Antall ansatte* falt ut kan ha sammenheng med at en eventuell effekt blir dominert av effekten av variablene *Nyregistrert* og *Enkeltmannsforetak* som viste seg betydningsfulle. Diskusjonen i avsnitt 6.3 viser at bortfallet også kan ha sammenheng med tilstedeværelsen av en enkelt ekstrem observasjon med endringsbeløp mer enn dobbelt så stor som den største av de andre beløpene i materialet.
- Den ekstreme observasjonen referert til i foregående punkt viste seg å ha en sterk innflytelse på estimert forventet endringsbeløp gitt avdekking, særlig for nyregistrerte enkteltmannsforetak uten ekstern regnskapsfører, men ikke på de estimerte avdekking-sannsynlighetene.
- Utvalget ble trukket i to trinn der det større utvalget på trinn ble gjenstand for en enklere og raskere undersøkelse enn full materiell kontroll, og ble brukt som screeningsgrunnlag for å øke sannsynligheten for å finne fram til endringstilfeller ved full og tidkrevende kontroll på det mindre trinn-2-utvalget. Det var to funn-indikatorer på trinn 1, *MAV* og *SV*, der *MAV* er basert på en skårvariabel regnet ut på grunnlag av revisors vurdering av kvaliteten av en rekke enkeltdimensjoner ved regnskap og bedriftsorganisering, og *SV* en verdi for revisors egen samlet vurdering av virksomheten. Bare *MAV* ble brukt som screeningsvariabel i denne undersøkelsen, men noe evidens i hovedrapporten tyder på at *SV* innholder noe tilleggsinformasjon om sannsynligheten for avdekking (inkludert "feil bruk av mva-satser"), slik at *SV* bør inngå i sreeningsgrunnlaget ved senere undersøkelser. Denne rapporten fant ytterligere støtte for denne konklusjonen. *SV* synes først og fremst å bidra i de tilfeller *MAV*-kriteriet ikke slår til, spesielt for nyregistrerte virksomheter. Det er også evidens for at *SV* bidrar til å oppdage "mindre alvorlige" endringstilfeller, målt med endringsbeløpets størrelse.

- Inkludering av data fra 2005 i materialet ga noe evidens for at det kan ha skjedd en utvikling av praksis i bruk av *SV*-kriteriet fra 2006 til 2007 (årene for innhenting av data). Når det gjelder *MAV* er det ingen tegn på det samme.
- Muligheten for regionale forskjeller er tonet ned i denne rapporten slik at forventet endringsbeløp og avdekkingsannsynligheter rapportert her kan oppfattes som gjennomsnittstørrelser over regionale forskjeller, som det ble funnet evidens for i hovedrapporten. Kommunetype (sentralitet og tjenesteyting) ble imidlertid beholdt blant kovariatene. Det viste seg at disse synes å ha en viss betydning for avdekking-sannsynligheten, med høyest sannsynlighet i ikke-sentrale kommuner som først og fremst er tjenesteytende, og minst sannsynlighet i de mest sentrale kommunene som ikke først og fremst er tjenesteytende. Når det gjaldt endringsbeløpene derimot, var det ingen tegn på sammenheng med kommunetype.

Appendiks A1. Matematiske oppdateringer

A1.1 Formler bak avsnitt 6

I dette avsnittet betyr variablene følgende:

Responsvariable:

$$
\begin{aligned}
& X = X_1 \\
(1) \qquad & Y = Y_1 \\
& Z_1 = Z_{mav} \\
& Z_2 = Z_{sv} ,
\end{aligned}
$$

eller kort

$$Z = (Z_1, Z_2) = (Z_{mav}, Z_{sv})$$

$Z = z$ betyr $Z_1 = z_1$ og $Z_2 = z_2$ der $z = (z_1, z_2)$

Eksogene kovariater:

(2) $U = (T, \text{ENK}, Nyreg, R, KSminKTJ)$

Uttrykket $E(X \mid U, Y, Z) = E(X \mid U, Y, Z_1, Z_2)$, som i utgangspunktet er en funksjon av U, Y, Z, skriver jeg noen ganger som $E_U(X \mid Y, Z)$ for å understreke at jeg ser på funksjonen partielt som en funksjon av Y, Z der U holdes fast, og med fordeling bestemt av den betingete fordelingen for $(Y, Z \mid U)$. For øvrig vil fordelingen under forventning og variansberegninger nedenfor avhenge av stratum *k* som undertrykkes i noen av uttrykkene. Argumentasjonen for øvrig følger stort sett av HRS der framstillingen er mer detaljert..

Skriv kort

$$(3)\qquad \begin{aligned} p_z &= p_{z_1 z_2} = P(Y=1 \mid U, Z=z) = P(Y=1 \mid U, Z_1=z_1, Z_2=z_2) \\ q_z &= q_{z_1 z_2} = P(Z=z \mid U) = P(Z_1=z_1, Z_2=z_2 \mid U) \end{aligned},$$

Setter vi dessuten

$$(4)\qquad q_1 = P(Z_1=1 \mid U) \quad \text{og} \quad q_2(z_1) = P(Z_2=1 \mid Z_1=z_1, U)$$

kan $q_z = q_{z_1 z_2}$ faktoriseres som

$$(5)\qquad q_z = q_{z_1 z_2} = q_1^{z_1}(1-q_1)^{1-z_1} q_2(z_1)^{z_2}(1-q_2(z_1))^{1-z_2}$$

Hvis V er en stokastisk variabel, betyr uttrykket $V \sim \Gamma(\mu, \varphi)$ at V er gammafordelt med forventning $E(V) = \mu$ og dispersjon φ, som innebærer at $\text{var}(V) = \varphi\mu^2$.

Modellen spesifiseres nå som

$$(6)\qquad \begin{aligned} & p_z = P(Y=1 \mid z, U) = \text{logit}^{-1}[(1, z, U)\beta] \\ & q_1 = P(Z_1=1 \mid U) = \text{logit}^{-1}[(1, U)\alpha] \\ & q_2(z_1) = P(Z_2=1 \mid z_1, U) = \text{logit}^{-1}[(1, z_1, U)\delta] \\ & P(X=0 \mid z, Y=0, U) = 1 \\ & (X \mid z, Y=1, U) \sim \Gamma(\mu_z, \varphi) \ \text{ der } \ \log(\mu_z) = [1, (1-z_1)z_2, U]\gamma \end{aligned}$$

der $\alpha, \delta, \beta, \gamma$ er variasjonsuavhengige parametervektorer. Videre antas at datamaterialet består av uavhengige og identisk fordelte observasjoner, med fordeling bestemt av (6), av vektoren, (*Z*,*U*), i hvert stratum på trinn 1 og likeledes for (*Z*,*Y*,*X*,*U*) i hvert stratum på trinn 2 med gitt Z_1. Utvalgsplanen innbærer mange manglende observasjoner av (*Y*,*X*), men, som påvist i appendiks A1.2 i HRS, har mangel-mekanismen MAR-karakter ("missing at random"), slik at den kan ignoreres i likelihood-funksjonen.

Vi finner prevalens-sannsynligheten $\overline{p} = P(Y=1 \mid U)$ gitt ved

$$(7)\qquad \overline{p} = E(Y \mid U) = E_U(Y) = E_U[E_U(Y \mid Z)] = E_U[p_Z] = \sum_z p_z q_z$$

der *z* gjennomløper parene (0,0), (0,1), (1,0), (1,1). Vi har dessuten

$$(8)\qquad E(X \mid U) = \sum_z \mu_z q_z p_z$$

$$(9)\qquad \text{var}(X \mid U) = (1+\varphi)\sum_z \mu_z^2 q_z p_z - \left[\sum_z \mu_z q_z p_z\right]^2$$

Bevis: Vi har

(i) $E_U(X \mid Y=y, Z=z) = \begin{cases} 0 & \text{med sanns.het} \quad 1-\overline{p} \quad \text{hvis } Y=0 \\ \mu_z & \text{med sanns.het} \quad q_z p_z \quad \text{hvis } Y=1 \end{cases}$

hvorav (8) følger. For (9) gjelder

$$\text{var}_U(X \mid Y=y, Z=z) = \begin{cases} 0 & \text{med sanns.het} \quad 1-\overline{p} \quad \text{hvis y} = 0 \\ \varphi\mu_z^2 & \text{med sanns.het} \quad q_z p_z \quad \text{hvis y} = 1 \end{cases}$$

som gir

$$E_U\left[\text{var}_U(X \mid Y, Z)\right] = \varphi \sum_z \mu_z^2 q_z p_z$$

hvorav, fra (i)

$$\begin{aligned} \text{var}(X \mid U) &= \text{var}_U\left[E_U(X \mid Y, Z)\right] + E_U\left[\text{var}_u(X \mid Y, Z)\right] \\ &= \sum_z \mu_z^2 q_z p_z - \left(\sum_z \mu_z q_z p_z\right)^2 + \varphi \sum_z \mu_z^2 q_z p_z \\ &= (1+\varphi)\sum_z \mu_z^2 q_z p_z - \left(\sum_z \mu_z q_z p_z\right)^2 \end{aligned}$$

Bevis slutt.

$E(X \mid U)$ omfatter blant annet virksomheter med $X = 0$. Av interesse vil derfor også være forventet endringsbeløp gitt avdekking, $E(X \mid U, Y=1)$ som blir

(10) $$E(X \mid U, Y=1) = E_U\left[E_U(X \mid Y=1, Z\right] = E_U \mu_Z = \sum_z \mu_z q_z$$

Sammenligner vi med (8), ser vi at $E(X \mid U) < E(X \mid U, Y=1)$ som forventet.

La U_i betegne observert verdi av *U* nr. *i* i utvalget fra stratum *k* (på trinn 1) der utvalgs- og stratumstørrelse er henholdsvis n_k og N_k. Estimater for $E_k(X)$ og $\text{var}_k(X)$ basert på utvalget er dermed gitt ved

$$\hat{E}_k(X) = \frac{1}{n_k}\sum_{i=1}^{n_k} E_k(X \mid U_i)$$

(11)

$$\hat{\text{var}}_k(X) = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(e_i - \bar{e})^2 + \bar{v} \quad \text{der} \quad e_i = E_k(X \mid U_i) \quad \text{og} \quad v_i = \text{var}_k(X \mid U_i)$$

Eksakte verdier, bortsett fra estimeringsusikkerhet, for $E_k(X)$ og $\text{var}_k(X)$ kan oppnås om SKD's database utnyttes til å finne U_i for alle enheter i stratumet.

Anta vi ønsker gjennomsnittlige verdier av $E_k(X)$ og $\text{var}_k(X)$ over en bransje som består av *K* strata. Kall gjennomsnittene for $\bar{E}(X)$ og $\overline{\text{var}}(X)$. Det er da rimelig å bruke gjennomsnitt som er veiet med stratum-størrelsen

$$\bar{E}(X) = \frac{1}{N}\sum_{k=1}^{K} N_k \hat{E}_k(X) \quad \text{der} \quad N = \sum_k N_k$$

(12)

$$\overline{\text{var}}(X) = \frac{1}{N}\sum_{k=1}^{K} N_k \hat{\text{var}}_k(X)$$

som sikrer at store strata veier mer i gjennomsnittet. Tilsvarende gjelder for gjennomsnitt av prevalens-sannsynligheten $\bar{p} = P(Y = 1 \mid U)$.

A1.2 Beregning av asymptotisk standardfeil for forventet endringsbeløp gitt avdekking-1 og sannsynligheten for avdekking-1.

Vi ønsker standardfeil for

(13) $$\hat{\bar{p}} = \hat{P}(Y = 1 \mid U) = \sum_z \hat{p}_z \hat{q}_z$$

og

(14) $$\hat{\bar{\mu}} = \hat{E}(X \mid U, Y = 1) = \sum_z \hat{\mu}_z \hat{q}_z$$

der $\bar{p}, \bar{\mu}$ er funksjoner av *U* og θ.

I henhold til (6) og modellen som beskrevet i avsnitt 6 er parametervektoren lik $\theta' = (\alpha', \delta', \beta', \gamma')$ og av dimensjon 14. La de estimerte kovariansmatrisene for de fire delene være $\Sigma_\alpha, \Sigma_\delta, \Sigma_\beta$ og Σ_γ henholdsvis. På grunn av faktoriseringen av likelihood-funksjonen (se HRS) blir (den asymptotiske) kovariansmatrisen for $\hat{\theta}$ blokk-diagonal:

$$\Sigma_\theta = \text{kovar}(\hat{\theta}) = \begin{pmatrix} \Sigma_\alpha & 0 & 0 & 0 \\ 0 & \Sigma_\delta & 0 & 0 \\ 0 & 0 & \Sigma_\beta & 0 \\ 0 & 0 & 0 & \Sigma_\gamma \end{pmatrix}$$

der 0-ene betegner 0-matriser av passende størrelse.

Som i HRS kan dispersjonsparameteren ignoreres.

Ifølge standard 1-ordens asymptotisk teori og Taylor-utvikling med ett ledd, blir asymptotisk varians for $\hat{\bar{p}}$:

$$\text{var}(\hat{\bar{p}}) = \frac{\partial \bar{p}}{\partial \theta'} \Sigma_\theta \frac{\partial \bar{p}}{\partial \theta} = \frac{\partial \bar{p}}{\partial \alpha'} \Sigma_\alpha \frac{\partial \bar{p}}{\partial \alpha} + \frac{\partial \bar{p}}{\partial \delta'} \Sigma_\delta \frac{\partial \bar{p}}{\partial \delta} + \frac{\partial \bar{p}}{\partial \beta'} \Sigma_\beta \frac{\partial \bar{p}}{\partial \beta} + \frac{\partial \bar{p}}{\partial \gamma'} \Sigma_\gamma \frac{\partial \bar{p}}{\partial \gamma}$$

og dermed asymptotisk standardfeil $\text{SE}(\hat{\bar{p}}) = \sqrt{\text{var}(\hat{\bar{p}})}$. (Tilsvarende for $\bar{\mu}$.)
Konsistente estimater fås ved å erstatte parametrene med ML-estimater.

Vi trenger derfor å utvikle uttrykk for de deriverte. Vi må da først presisere uttrykkene i modellen i (6). La u = (*T*, ENK, *Nyreg*, *R*, *KsminKtj*)' være vektoren av kovariater i modellen, og la z_1, z_2 som ovenfor stå for Z_{mav}, Z_{sv}.

Videre utvider vi kovariat-vektoren med konstantleddet ved

$$u_1' = (u', 1)$$

Så lager vi utplukks-matriser som plukker ut de variable fra kovariat-matrisene som inngår. For eksempel logit-uttrykket for $q_2(z_1)$, som har kovariatene, z_1,*T*, *Nyreg*, *konstant*, med i modellen, kan nå skrives

$$q_2(z_1) = \text{logit}^{-1}(\delta'(D_{z\delta} z + D_\delta u_1)$$

der

$$D_{z\delta} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{og} \quad D_\delta = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Den deriverte av $q_2(z_1)$ blir

$$\frac{\partial}{\partial\delta}q_2(z_1) = q_2(z_1)\cdot(1-q_2(z_1))\cdot(D_{z\delta}z + D_\delta u_1)$$

og tilsvarende for de andre uttrykkene i (6).

Uttrykket for q_z i (5) kan forenkles litt ved å utnytte at hvis *q* er et tall og w er 0 eller 1, så er

$$q^w(1-q)^w = wq + (1-q)(1-w) = (2w-1)q + 1$$

Litt algebra gir (*z* betyr kolonnevektoren $(z_1, z_2)'$ i formlene nedenfor):

$$\frac{\partial q_z}{\partial\alpha} = (2z_1 - 1)q_1^{1-z_1}(1-q_1)^{z_1}q_z(D_{z\alpha}z + D_\alpha u_1)$$

$$\frac{\partial q_z}{\partial\delta} = (2z_2 - 1)q_2(z_1)^{1-z_2}(1-q_2(z_1))^{z_2}q_z(D_{z\delta}z + D_\delta u_1)$$

$$\frac{\partial p_z}{\partial\beta} = p_z(1-p_z)\cdot(D_{z\beta}z + D_\beta u_1)$$

$$\frac{\partial\mu_z}{\partial\gamma} = \mu_z(D_{z\gamma}\tilde{z} + D_\gamma u_1) \qquad \text{der} \quad \tilde{z} = [z_1, (1-z_1)z_2]'$$

Av (13) og (14) får vi dermed

$$\frac{\partial\overline{p}}{\partial\alpha} = \sum_z p_z q_z (2z_1 - 1)\, q_1^{1-z_1}(1-q_1)^{z_1}\cdot(D_{z\alpha}z + D_\alpha u_1)$$

$$\frac{\partial\overline{p}}{\partial\delta} = \sum_z p_z q_z (2z_2 - 1)\, q_2(z_1)^{1-z_2}(1-q_2(z_1))^{z_2}\cdot(D_{z\delta}z + D_\delta u_1)$$

$$\frac{\partial\overline{p}}{\partial\beta} = \sum_z q_z p_z(1-p_z)\cdot(D_{z\beta}z + D_\beta u_1)$$

$$\frac{\partial\overline{p}}{\partial\gamma} = 0\cdot D_\gamma u_1$$

og

$$\frac{\partial \bar{\mu}}{\partial \alpha} = \sum_z \mu_z q_z (2z_1 - 1) q_1^{1-z_1} (1-q_1)^{z_1} \cdot (D_{z\alpha} z + D_\alpha u_1)$$

$$\frac{\partial \bar{\mu}}{\partial \delta} = \sum_z \mu_z q_z (2z_2 - 1)\, q_2(z_1)^{1-z_2} (1-q_2(z_1))^{z_2} \cdot (D_{z\delta} z + D_\delta u_1)$$

$$\frac{\partial \bar{\mu}}{\partial \beta} = 0 \cdot D_\beta u_1$$

$$\frac{\partial \bar{\mu}}{\partial \gamma} = \sum_z q_z \mu_z (D_{z\gamma} \tilde{z} + D_\gamma u_1) \qquad \text{der} \quad \tilde{z} = [z_1, (1-z_1) z_2]'$$

Appendiks A2 Utskrifter

(Hovedsakelig basert på STATA)

A2.1 Utskrifter for Y_1 (indikator for avdekking av typen "sum-aarsak")

Ut1 Y1: Full modell - Alternativ 1: hele materialet

```
Logistic regression                                     Number of obs   =        135
                                                        LR chi2(11)     =      33.51
                                                        Prob > chi2     =     0.0004
Log likelihood = -46.423624                             Pseudo R2       =     0.2652

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.567528   .6337064     2.47   0.013     .3254863     2.80957
        Zmav |    .3668886   .8656173     0.42   0.672     -1.32969    2.063467
         Zsv |    .6204736   .7164038     0.87   0.386     -.783652    2.024599
      Engros |    2.558017   1.179259     2.17   0.030     .2467118    4.869322
       Rengj |    2.366882   1.128598     2.10   0.036     .1548695    4.578894
       Nyreg |     .827447    .631496     1.31   0.190    -.4102624    2.065156
         ENK |    .6446357   .7016256     0.92   0.358    -.7305253    2.019797
    Ansatte0 |    .1980753   .6493441     0.31   0.760    -1.074616    1.470766
           R |   -.5859422   .6204228    -0.94   0.345    -1.801949    .6300642
  Komsentral |   -.8105018    .612399    -1.32   0.186    -2.010782    .3897782
  KomTjenest |    1.173607   .6209414     1.89   0.059    -.0434163    2.390629
       _cons |   -5.294039   1.425777    -3.71   0.000    -8.088512   -2.499567
------------------------------------------------------------------------------
```

Ut2 Y1: Redusert modell 1 - Alternativ 1: hele materialet

```
Logistic regression                                     Number of obs   =        135
                                                        LR chi2(6)      =      27.95
                                                        Prob > chi2     =     0.0001
Log likelihood = -49.204718                             Pseudo R2       =     0.2212

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.743412   .5835447     2.99   0.003     .5996853    2.887139
        Zmav |    1.190991    .592458     2.01   0.044     .0297947    2.352188
      Engros |    2.896329   1.137694     2.55   0.011     .6664908    5.126168
       Rengj |    2.759069   1.109111     2.49   0.013     .5852528    4.932886
  Komsentral |   -.8573825   .5810139    -1.48   0.140    -1.996149    .2813839
  KomTjenest |    .9763902   .5780752     1.69   0.091    -.1566163    2.109397
       _cons |   -5.331282   1.221848    -4.36   0.000     -7.72606   -2.936504
------------------------------------------------------------------------------
```

Ut3 Y1: Redusert modell 2 - Alternativ 1: hele materialet. Engros og Rengj slått sammen.

```
Logistic regression                               Number of obs   =        135
                                                  LR chi2(5)      =      27.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -49.235101                       Pseudo R2       =     0.2207

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.753715   .5816447     3.02   0.003      .613712    2.893717
        Zmav |    1.173283   .5895412     1.99   0.047     .0178034    2.328763
 EngrosRengj |    2.812959   1.087097     2.59   0.010     .6822875     4.94363
  Komsentral |   -.8678752   .5808034    -1.49   0.135    -2.006229    .2704786
  KomTjenest |    .9814455   .5785873     1.70   0.090    -.1525648    2.115456
       _cons |   -5.337222   1.220398    -4.37   0.000    -7.729159   -2.945285
------------------------------------------------------------------------------
```

Ut4 Y1: Redusert modell 3 - Alternativ 1: hele materialet. Engros og Rengj slått sammen og KSminusKTJ erstatter Komsentr og Komtj.

```
Logistic regression                               Number of obs   =        135
                                                  LR chi2(4)      =      27.86
                                                  Prob > chi2     =     0.0000
Log likelihood = -49.251754                       Pseudo R2       =     0.2205

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.766681   .5778804     3.06   0.002     .6340559    2.899305
        Zmav |    1.185712   .5850113     2.03   0.043     .0391111    2.332313
 EngrosRengj |    2.804832   1.086224     2.58   0.010     .6758719    4.933792
  KSminusKTJ |   -.9251269    .490539    -1.89   0.059    -1.886566    .0363119
       _cons |   -5.296386   1.198732    -4.42   0.000    -7.645858   -2.946915
------------------------------------------------------------------------------
```

Ut5 Y1: Redusert modell 4 - Alternativ 1: hele materialet. Engros og Rengj slått sammen og uten Zmav og KSminusKTJ.

```
Logistic regression                               Number of obs   =        137
                                                  LR chi2(2)      =      20.36
                                                  Prob > chi2     =     0.0000
Log likelihood = -53.389162                       Pseudo R2       =     0.1601

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.738767   .5551773     3.13   0.002     .6506395    2.826894
 EngrosRengj |    2.952892    1.05741     2.79   0.005     .8804061    5.025378
       _cons |   -5.235275   1.156946    -4.53   0.000    -7.502847   -2.967702
------------------------------------------------------------------------------
```

Ut6 Y1: Full modell - Alternativ 2: Bare for Engros+Rengjøring data

```
Logistic regression                                     Number of obs   =        102
                                                        LR chi2(10)     =      24.56
                                                        Prob > chi2     =     0.0062
Log likelihood = -42.163834                             Pseudo R2       =     0.2256

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.589262   .6370283     2.49   0.013     .3407091    2.837814
        Zmav |    .5719661   .9232762     0.62   0.536    -1.237622    2.381554
         Zsv |    .5969162   .7539043     0.79   0.428     -.880709    2.074541
      Engros |    .1807805   .6432566     0.28   0.779    -1.079979     1.44154
       Nyreg |    .8505423   .6447252     1.32   0.187    -.4130959    2.114181
         ENK |    .5187999   .7237856     0.72   0.474    -.8997938    1.937394
    Ansatte0 |     .078081   .6786986     0.12   0.908    -1.252144    1.408306
           R |    -.637346   .6262384    -1.02   0.309    -1.864751    .5900588
  Komsentral |    -.994809   .6530207    -1.52   0.128    -2.274706     .285088
  KomTjenest |    1.065835    .646662     1.65   0.099    -.2015995    2.333269
       _cons |   -2.690717   .8431967    -3.19   0.001    -4.343352   -1.038082
------------------------------------------------------------------------------
```

Ut7 Y1: Redusert modell 1 - Alternativ 2: Bare for Engros+Rengjøring data

```
Logistic regression                                     Number of obs   =        102
                                                        LR chi2(3)      =      19.55
                                                        Prob > chi2     =     0.0002
Log likelihood =  -44.67195                             Pseudo R2       =     0.1795

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.778159   .5822712     3.05   0.002     .6369286     2.91939
        Zmav |    1.267215   .6050662     2.09   0.036     .0813067    2.453123
  KSminusKTJ |   -.9910238    .524264    -1.89   0.059    -2.018562    .0365147
       _cons |    -2.51926   .5242885    -4.81   0.000    -3.546847   -1.491673
------------------------------------------------------------------------------
```

Ut8 Y1: Redusert modell 2 - Alternativ 2: Bare for Engros+Rengjøring data. Uten Z_{mav} og KSminusKTJ

```
Logistic regression                                     Number of obs   =        103
                                                        LR chi2(1)      =      11.64
                                                        Prob > chi2     =     0.0006
Log likelihood = -48.877654                             Pseudo R2       =     0.1064

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |    1.738767   .5551772     3.13   0.002     .6506395    2.826894
       _cons |   -2.282382   .4694765    -4.86   0.000    -3.202539   -1.362225
------------------------------------------------------------------------------
```

A2.2 Utskrifter for Y (indikator for avdekking av typen "endringer alle typer")

Ut9 Y: Full modell - Alternativ 1

```
Logistic regression                                     Number of obs   =        135
                                                        LR chi2(11)     =      29.01
                                                        Prob > chi2     =     0.0023
Log likelihood = -55.729446                             Pseudo R2       =     0.2065

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.269865   .5541145     2.29   0.022     .1838202    2.355909
        Zmav |   .6874652   .7752933     0.89   0.375    -.8320818    2.207012
         Zsv |   .6423341   .6466179     0.99   0.321    -.6250136    1.909682
      Engros |   1.909321   .9143635     2.09   0.037     .1172017    3.701441
       Rengj |   1.765763   .8601494     2.05   0.040     .0799014    3.451625
       Nyreg |   .8673788   .5703544     1.52   0.128    -.2504952    1.985253
         ENK |   .3644761    .630044     0.58   0.563    -.8703875     1.59934
    Ansatte0 |  -.5151567   .5973118    -0.86   0.388    -1.685866    .6555529
           R |  -.1447954   .5725908    -0.25   0.800    -1.267053     .977462
  Komsentral |  -.6192947   .5386085    -1.15   0.250    -1.674948    .4363586
  KomTjenest |   .5381355   .5594043     0.96   0.336    -.5582767    1.634548
       _cons |  -3.921291   1.112145    -3.53   0.000    -6.101055   -1.741527
------------------------------------------------------------------------------
```

Ut10 Y: Redusert modell 1 - Alternativ 1 (med EngrosRengj)
(P-verdi vs full modell: 0.828, df = 7)

```
Logistic regression                                     Number of obs   =        135
                                                        LR chi2(4)      =      25.45
                                                        Prob > chi2     =     0.0000
Log likelihood = -57.512776                             Pseudo R2       =     0.1811

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.363365   .5207205     2.62   0.009     .3427717    2.383959
        Zmav |   1.043899   .5698112     1.83   0.067    -.0729104    2.160709
 EngrosRengj |   1.980001   .8084347     2.45   0.014     .3954977    3.564504
       Nyreg |   .9469858   .5134606     1.84   0.065    -.0593785     1.95335
       _cons |  -4.274225   .8979172    -4.76   0.000     -6.03411   -2.514339
------------------------------------------------------------------------------
```

Ut11 Y: Redusert modell 2 - Alternativ 1 (med EngrosRengj)
(P-verdi vs full modell: 0.232, df = 9)

```
Logistic regression                                     Number of obs   =        137
                                                        LR chi2(2)      =      18.30
                                                        Prob > chi2     =     0.0001
Log likelihood = -61.566215                             Pseudo R2       =     0.1294

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.532674   .4986146     3.07   0.002     .5554073    2.509941
 EngrosRengj |   2.401025   .7846834     3.06   0.002     .8630739    3.938976
       _cons |  -4.305263   .8830979    -4.88   0.000    -6.036103   -2.574423
------------------------------------------------------------------------------
```

Ut12 Y: Redusert modell 1 - Alternativ 2: (med Zmav)

```
Logistic regression                                     Number of obs   =        102
                                                        LR chi2(3)      =      18.71
                                                        Prob > chi2     =     0.0003
Log likelihood =  -49.59365                             Pseudo R2       =     0.1587

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.350094   .5231403     2.58   0.010     .3247579     2.37543
        Zmav |   1.092442   .5834893     1.87   0.061     -.051176     2.23606
       Nyreg |   1.053058   .5337518     1.97   0.049     .0069237    2.099192
       _cons |  -2.332171    .461453    -5.05   0.000    -3.236602    -1.42774
------------------------------------------------------------------------------
```

Ut13 Y: Redusert modell 1 - Alternativ 2: (med Zsv)

```
Logistic regression                                     Number of obs   =        103
                                                        LR chi2(3)      =      19.24
                                                        Prob > chi2     =     0.0002
Log likelihood = -49.631373                             Pseudo R2       =     0.1624

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.176966   .5282292     2.23   0.026     .1416561    2.212277
         Zsv |   1.063704   .5171354     2.06   0.040     .0501376    2.077271
       Nyreg |   .8694435   .5496347     1.58   0.114    -.2078207    1.946708
       _cons |  -2.348949   .4558833    -5.15   0.000    -3.242464   -1.455434
------------------------------------------------------------------------------
```

Ut14 Y: Redusert modell 1 - Alternativ 2: (med både Zmav og Zsv)

```
Logistic regression                                     Number of obs   =        102
                                                        LR chi2(4)      =      19.93
                                                        Prob > chi2     =     0.0005
Log likelihood = -48.984212                             Pseudo R2       =     0.1690

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.246855   .5345307     2.33   0.020     .1991938    2.294516
        Zmav |   .6188842   .7176152     0.86   0.388    -.7876158    2.025384
         Zsv |   .7149109   .6415866     1.11   0.265    -.5425758    1.972398
       Nyreg |   .9026422   .5570344     1.62   0.105    -.1891251     1.99441
       _cons |   -2.39248   .4664578    -5.13   0.000     -3.30672    -1.47824
------------------------------------------------------------------------------
```

Ut15 Y: Redusert modell 1 - Alternativ 2: (med både Zmav og Zaltsv)

```
Logistic regression                                     Number of obs   =        102
                                                        LR chi2(4)      =      19.15
                                                        Prob > chi2     =     0.0007
Log likelihood = -49.375108                             Pseudo R2       =     0.1624

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.278671   .5351736     2.39   0.017       .22975    2.327592
        Zmav |    1.19995   .6035934     1.99   0.047     .0169283    2.382971
      Zaltsv |   .4635221   .6937038     0.67   0.504    -.8961123    1.823157
       Nyreg |   .9686493   .5508416     1.76   0.079    -.1109804    2.048279
       _cons |    -2.3701   .4655183    -5.09   0.000    -3.282499   -1.457701
------------------------------------------------------------------------------
```

A2.3 Utskrifter for X_1 (endringsbeløp av typen "sum-aarsak")

Alternativ 1:

****

Ut16 X1: Full modell - Alternativ 1 (med Zsv)

```
Generalized linear models                         No. of obs      =         24
Optimization     : ML                             Residual df     =         12
                                                  Scale parameter =   1.051539
Deviance         =  18.54178813                   (1/df) Deviance =   1.545149
Pearson          =  12.61846247                   (1/df) Pearson  =   1.051539

Variance function: V(u) = u^2                     [Gamma]
Link function    : g(u) = ln(u)                   [Log]
                                                  AIC             =   27.04528
Log likelihood   = -312.5433507                   BIC             =  -19.59486

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -.3157489   .6533466    -0.48   0.629    -1.596285     .964787
         Zsv |  -1.257269   .8644171    -1.45   0.146    -2.951495    .4369575
        Zmav |   1.602874   .8756762     1.83   0.067    -.1134197    3.319168
      Engros |   -2.82247   1.535099    -1.84   0.066    -5.831208    .1862684
       Rengj |  -2.586975   1.425447    -1.81   0.070    -5.380799    .2068492
       Nyreg |   .8690774   .6581625     1.32   0.187    -.4208974    2.159052
         ENK |  -.4492552   .7213207    -0.62   0.533    -1.863018    .9645074
    Ansatte0 |  -.5523592   .6839652    -0.81   0.419    -1.892906    .7881878
           R |   -1.05005   .5505483    -1.91   0.056    -2.129105    .0290049
  Komsentral |  -.4503608   .5555984    -0.81   0.418    -1.539314     .638592
  KomTjenest |  -.4243631   .6387167    -0.66   0.506    -1.676225    .8274986
       _cons |   16.34182   2.171022     7.53   0.000     12.08669    20.59694
------------------------------------------------------------------------------
```

Ut17 X1: Redusert modell 1 - Alternativ 1 (med Zsv)

```
Generalized linear models                         No. of obs      =         24
Optimization     : ML                             Residual df     =         19
                                                  Scale parameter =   .8177796
Deviance         =  23.12524631                   (1/df) Deviance =   1.217118
Pearson          =  15.53781173                   (1/df) Pearson  =   .8177796

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

                                                  AIC             =   26.65292
Log likelihood   = -314.8350798                   BIC             =  -37.25778
------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Nyreg |   .9353122   .4849068     1.93   0.054    -.0150877    1.885712
         Zsv |  -1.539536   .6134617    -2.51   0.012    -2.741899   -.3371729
        Zmav |   1.004037    .515814     1.95   0.052      -.00694    2.015014
           R |  -.8002759   .4348362    -1.84   0.066    -1.652539    .0519874
       _cons |   12.76117   .4659168    27.39   0.000     11.84799    13.67435
------------------------------------------------------------------------------
```

Ut18 X1: Redusert modell 2 - Alternativ 1

```
Generalized linear models                         No. of obs      =         24
Optimization     : ML                             Residual df     =         23
                                                  Scale parameter =   1.330093
Deviance         =  31.33659542                   (1/df) Deviance =   1.362461
Pearson          =  30.59214755                   (1/df) Pearson  =   1.330093

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

                                                  AIC             =  26.66173
Log likelihood   = -318.9407543                   BIC             = -41.75864

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |    12.2892   .2354313    52.20   0.000     11.82776    12.75063
------------------------------------------------------------------------------
```

Ut19 X1: Full modell - Alternativ 1 (med Zaltsv istedenfor Zsv)

```
Generalized linear models                         No. of obs      =         24
Optimization     : ML                             Residual df     =         12
                                                  Scale parameter =   1.051539
Deviance         =  18.54178813                   (1/df) Deviance =   1.545149
Pearson          =  12.61846247                   (1/df) Pearson  =   1.051539

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

                                                  AIC             =  27.04528
Log likelihood   = -312.5433507                   BIC             = -19.59486

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -.3157489   .6533466    -0.48   0.629    -1.596285     .964787
        Zmav |   .3456052   .9228688     0.37   0.708    -1.463184    2.154395
      Zaltsv |  -1.257269   .8644171    -1.45   0.146    -2.951495    .4369575
      Engros |   -2.82247   1.535099    -1.84   0.066    -5.831208    .1862684
       Rengj |  -2.586975   1.425447    -1.81   0.070    -5.380799    .2068492
       Nyreg |   .8690774   .6581625     1.32   0.187    -.4208974    2.159052
         ENK |  -.4492552   .7213207    -0.62   0.533    -1.863018    .9645074
    Ansatte0 |  -.5523592   .6839652    -0.81   0.419    -1.892906    .7881878
           R |   -1.05005   .5505483    -1.91   0.056    -2.129105    .0290049
  Komsentral |  -.4503608   .5555984    -0.81   0.418    -1.539314     .638592
  KomTjenest |  -.4243631   .6387167    -0.66   0.506    -1.676225    .8274986
       _cons |   16.34182   2.171022     7.53   0.000     12.08669    20.59694
------------------------------------------------------------------------------
```

Ut20 X1: Redusert modell 3 - Alternativ 1 (med Zaltsv istedenfor Zsv)

```
Generalized linear models                         No. of obs      =         24
Optimization     : ML                             Residual df     =         21
                                                  Scale parameter =   .795785
Deviance         =  26.03095989                   (1/df) Deviance =   1.23957
Pearson          =  16.71148547                   (1/df) Pearson  =   .795785

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

                                                  AIC             =  26.60733
Log likelihood   = -316.2879366                   BIC             = -40.70817

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Zaltsv |  -1.039616   .4639708    -2.24   0.025    -1.948982   -.1302503
       Nyreg |   .7402161    .382198     1.94   0.053    -.0088783     1.48931
       _cons |   12.08683   .2449405    49.35   0.000     11.60675     12.5669
------------------------------------------------------------------------------
```

Alternativ 2:

****

Ut21 X1: Full modell - Alternativ 2 (med Zsv)

```
Generalized linear models                         No. of obs      =         23
Optimization     : ML                             Residual df     =         12
                                                  Scale parameter =   1.051539
Deviance         =  18.54178813                   (1/df) Deviance =   1.545149
Pearson          =  12.61846247                   (1/df) Pearson  =   1.051539

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

                                                  AIC             =  26.90815
Log likelihood   = -298.4436727                   BIC             = -19.08414

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -.3157489   .6533467    -0.48   0.629    -1.596285     .964787
        Zmav |   1.602874   .8756762     1.83   0.067    -.1134198    3.319168
         Zsv |  -1.257269   .8644172    -1.45   0.146    -2.951495    .4369576
      Engros |   -.235495   .5932711    -0.40   0.691    -1.398285     .927295
       Nyreg |   .8690774   .6581625     1.32   0.187    -.4208974    2.159052
         ENK |  -.4492552   .7213207    -0.62   0.533    -1.863018    .9645075
    Ansatte0 |  -.5523592   .6839652    -0.81   0.419    -1.892906    .7881879
  Komsentral |  -.4503608   .5555984    -0.81   0.418    -1.539314    .6385921
  KomTjenest |  -.4243631   .6387167    -0.66   0.506    -1.676225    .8274987
           R |   -1.05005   .5505484    -1.91   0.056    -2.129105    .0290049
       _cons |   13.75484   1.081948    12.71   0.000     11.63426    15.87542
------------------------------------------------------------------------------
```

Ut22 X1: Redusert modell 1 - Alternativ 2 (med Zsv)

```
Generalized linear models                         No. of obs      =         23
Optimization     : ML                             Residual df     =         18
                                                  Scale parameter =  .7331659
Deviance         =  20.57777297                   (1/df) Deviance =   1.14321
Pearson          =   13.1969866                   (1/df) Pearson  =  .7331659

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

                                                  AIC             =  26.47493
Log likelihood   = -299.4616651                   BIC             = -35.86112

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Nyreg |   1.016366   .4473906     2.27   0.023     .1394962    1.893235
        Zmav |   1.008182   .4886195     2.06   0.039     .0505051    1.965858
         Zsv |  -1.381005   .5682442    -2.43   0.015    -2.494743   -.2672664
           R |  -.8905276   .4052821    -2.20   0.028    -1.684866   -.0961892
       _cons |   12.58885   .4316053    29.17   0.000     11.74292    13.43478
------------------------------------------------------------------------------
```

Ut23 X1: Redusert modell 2 - Alternativ 2

```
Generalized linear models                         No. of obs      =         23
Optimization     : ML                             Residual df     =         22
                                                  Scale parameter =  1.472102
Deviance         =  30.38919648                   (1/df) Deviance =  1.381327
Pearson          =   32.3862414                   (1/df) Pearson  =  1.472102

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

                                                  AIC             =  26.55368
Log likelihood   = -304.3673769                   BIC             = -38.59168

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   12.23336    .253013    48.35   0.000     11.73747    12.72926
------------------------------------------------------------------------------
```

Ut24 X1: Full modell - Alternativ 2 (med Zaltsv istedenfor Zsv)

```
Generalized linear models                         No. of obs      =         23
Optimization     : ML                             Residual df     =         12
                                                  Scale parameter =   1.051539
Deviance         =  18.54178813                   (1/df) Deviance =   1.545149
Pearson          =  12.61846247                   (1/df) Pearson  =   1.051539

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

                                                  AIC             =  26.90815
Log likelihood   = -298.4436727                   BIC             = -19.08414

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -.3157489   .6533467    -0.48   0.629    -1.596285     .964787
        Zmav |   .3456052   .9228689     0.37   0.708    -1.463185    2.154395
      Zaltsv |  -1.257269   .8644172    -1.45   0.146    -2.951495    .4369576
      Engros |   -.235495   .5932711    -0.40   0.691    -1.398285     .927295
       Nyreg |   .8690774   .6581625     1.32   0.187    -.4208974    2.159052
         ENK |  -.4492552   .7213207    -0.62   0.533    -1.863018    .9645075
    Ansatte0 |  -.5523592   .6839652    -0.81   0.419    -1.892906    .7881879
           R |   -1.05005   .5505484    -1.91   0.056    -2.129105    .0290049
  Komsentral |  -.4503608   .5555984    -0.81   0.418    -1.539314    .6385921
  KomTjenest |  -.4243631   .6387167    -0.66   0.506    -1.676225    .8274987
       _cons |   13.75484   1.081948    12.71   0.000     11.63426    15.87542
------------------------------------------------------------------------------
```

Ut25 X1: Redusert modell 3 - Alternativ 2 (med Zaltsv istedenfor Zsv)

```
Generalized linear models                         No. of obs      =         23
Optimization     : ML                             Residual df     =         19
                                                  Scale parameter =    .72709
Deviance         =  20.96466442                   (1/df) Deviance =   1.103403
Pearson          =  13.81470994                   (1/df) Pearson  =    .72709

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

                                                  AIC             =  26.40479
Log likelihood   = -299.6551108                   BIC             = -38.60973

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           R |    -.79005   .3805896    -2.08   0.038    -1.535992   -.0441081
       Nyreg |   .8355247   .3769562     2.22   0.027     .0967041    1.574345
      Zaltsv |  -1.140413   .4600158    -2.48   0.013    -2.042028   -.2387991
       _cons |   12.42838   .3623281    34.30   0.000     11.71823    13.13853
------------------------------------------------------------------------------
```

A2.4 Utskrifter for Z_{mav} og Z_{sv} (funn på trinn 1)

Alternativ 2:

****

Ut26 Zsv|Zmav: Full modell - Alternativ 2

```
Logistic regression                               Number of obs   =        293
                                                  LR chi2(9)      =      81.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -94.605613                       Pseudo R2       =     0.3015

------------------------------------------------------------------------------
         Zsv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.133605   .4654237     2.44   0.015     .2213909    2.045818
        Zmav |   4.148859   .7213359     5.75   0.000     2.735066    5.562651
      Engros |    .264667   .4374384     0.61   0.545    -.5926964     1.12203
       Nyreg |    1.12827   .4098172     2.75   0.006     .3250434    1.931497
         ENK |   .3701694   .4852122     0.76   0.446     -.580829    1.321168
    Ansatte0 |   .3126001   .4379086     0.71   0.475    -.5456849    1.170885
           R |  -.0393762   .4586788    -0.09   0.932    -.9383701    .8596178
  Komsentral |    .002722   .4088092     0.01   0.995    -.7985294    .8039734
  KomTjenest |   .0539054    .427569     0.13   0.900    -.7841145    .8919252
       _cons |  -3.611214   .6128733    -5.89   0.000    -4.812424   -2.410004
------------------------------------------------------------------------------
```

Ut27 Zsv|Zmav: Redusert modell - Alternativ 2

```
Logistic regression                               Number of obs   =        293
                                                  LR chi2(3)      =      79.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -95.704851                       Pseudo R2       =     0.2934

------------------------------------------------------------------------------
         Zsv |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |   1.211507    .449335     2.70   0.007      .330827    2.092188
        Zmav |   4.360006   .7102997     6.14   0.000     2.967844    5.752168
       Nyreg |   1.146554   .3916082     2.93   0.003     .3790158    1.914092
       _cons |  -3.260775   .4299897    -7.58   0.000     -4.10354   -2.418011
------------------------------------------------------------------------------
```

Ut28 Zmav: Full modell - Alternativ 2

```
Logistic regression                               Number of obs   =        293
                                                  LR chi2(8)      =      24.45
                                                  Prob > chi2     =     0.0019
Log likelihood = -68.373214                       Pseudo R2       =     0.1517

------------------------------------------------------------------------------
        Zmav |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           T |  -.0764335   .4950215    -0.15   0.877    -1.046658    .8937909
      Engros |   .1428854   .5442433     0.26   0.793    -.9238119    1.209583
       Nyreg |   .4288801   .5202078     0.82   0.410    -.5907083    1.448469
         ENK |   2.523839   .7243506     3.48   0.000     1.104138     3.94354
    Ansatte0 |   .1660929   .5304368     0.31   0.754    -.8735442     1.20573
           R |  -.5796618   .5208854    -1.11   0.266    -1.600578    .4412548
  Komsentral |   .1426403   .4870816     0.29   0.770     -.812022    1.097303
  KomTjenest |    .141024   .5136282     0.27   0.784    -.8656688    1.147717
       _cons |  -4.034254   .7941622    -5.08   0.000    -5.590783   -2.477725
------------------------------------------------------------------------------
```

Ut29 Zmav: Redusert modell - Alternativ 2

```
Logistic regression                                     Number of obs   =        293
                                                        LR chi2(1)      =      22.14
                                                        Prob > chi2     =     0.0000
Log likelihood = -69.531378                             Pseudo R2       =     0.1373

------------------------------------------------------------------------------
        Zmav |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ENK |   2.396111   .6316972     3.79   0.000     1.158007    3.634215
       _cons |  -4.025352   .5824409    -6.91   0.000    -5.166915   -2.883788
------------------------------------------------------------------------------
```

A2.5 Utskrifter for tabell 6.5

Ut30 Y1: forenklet modell i tabell 6.5 - Alternativ 2

```
Logistic regression                                     Number of obs   =        102
                                                        LR chi2(2)      =      15.65
                                                        Prob > chi2     =     0.0004
Log likelihood = -46.617494                             Pseudo R2       =     0.1438

------------------------------------------------------------------------------
          Y1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Zmav |   1.122659   .5865045     1.91   0.056    -.0268682    2.272187
           T |   1.736651   .5657464     3.07   0.002     .6278082    2.845493
       _cons |  -2.522577   .5037771    -5.01   0.000    -3.509962   -1.535191
------------------------------------------------------------------------------
```

Ut31 X1: forenklet modell i tabell 6.5 - Alternativ 2

```
Generalized linear models                           No. of obs      =         23
Optimization     : ML                               Residual df     =         22
                                                    Scale parameter =   1.472102
Deviance         =  30.38919648                     (1/df) Deviance =   1.381327
Pearson          =   32.3862414                     (1/df) Pearson  =   1.472102

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

                                                    AIC             =   26.55368
Log likelihood   = -304.3673769                     BIC             =  -38.59168

------------------------------------------------------------------------------
             |                 OIM
           X |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   12.23336    .253013    48.35   0.000     11.73747    12.72926
------------------------------------------------------------------------------
```

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

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

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

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