Fraud detection by a multinomial model: Separating honesty from unobserved fraud

BY Jonas Andersson, Andreas Olden and Aija Rusina

DISCUSSION PAPER

NHH

Institutt for foretaksøkonomi
Department of Business and Management Science

FOR 15/2020

ISSN: 1500-4066
December 2020

Fraud detection by a multinomial model: Separating honesty from unobserved fraud

Jonas Andersson [*], Andreas Olden[†], and Aija Rusina[‡]

Norwegian School of Economics, Department of Business and Management Science and NoCeT

December 31, 2020

Abstract

In this paper we investigate the EM-estimator of the model by Caudill et al. (2005). The purpose of the model is to identify items, e.g. individuals or companies, that are wrongly classified as honest; an example of this is the detection of tax evasion. Normally, we observe two groups of items, labeled *fradulent* and *honest*, but suspect that many of the observationally honest items are, in fact, fraudulent. The items observed as *honest* are therefore divided into two unobserved groups, *honestH*, representing the truly honest, and *honestF*, representing the items that are observed as honest, but that are actually fraudulent. By using a multinomial logit model and assuming commonality between the observed *fradulent* and the unobserved *honestF*, Caudill et al. (2005) present a method that uses the EM-algorithm to separate them. By means of a Monte Carlo study, we investigate how well the method performs, and under what circumstances. We also study how well boostrapped standard errors estimates the standard deviation of the parameter estimators.

1 Introduction

Fraud is a fact of social behaviour having increasingly important consequences including loss of revenues to businesses, government, and society. Fraud is also expensive, driving up cost for detection and fraud risk reduction. As a result, active fraud control has gradually become an integrated part of business decision-making processes. Insurance companies must deal with fraud perpetrated by consumers on the firm and spend money on fraud detection and monitoring. A lot of research has focused on the fraud detection efforts and the frequency

[*] jonas.andersson@nhh.no

[†] andreasolden@gmail.com

[‡] aija.rusina.polakova@gmail.com

of fraud, that is, assessing and ranking the fraud suspiciousness of individual claims (Ai et al., 2009, 2013; Artís et al., 1999; Brockett et al., 2002; Derrig and Ostaszewski, 1995; Viaene et al., 2002).

Fraud is often detected by some sort of audit process. Audits will normally reveal some information about the fraudsters, the type of situations where fraud occurs, or the products where fraud is common. This information is then used to predict which claims are more likely to be fraudulent in the future. However, for most audits, the detection probability is not one hundred percent, which skews the estimated probabilities. This occurs because there are people in the group assumed not to be fraudulent who are actually fraudulent, making the observed fraudulent and the observed honest too similar. We expect that much data has this structure, in particular insurance claims data, tax data, and medical or diagnostic data. Moreover, the larger the fraction of misclassified observations, the worse the problem becomes.

Numerous studies develop techniques to identify or classify fraudulent claims. Predictive techniques are used to predict values for a certain target variable, such as credit scoring to predict repayment behaviour of loan applicants, and logistic regression models, both binary and multinomial logit models, are used for detecting manipulation such as dishonest insurance claims (Major and Riedinger, 2002; Olinsky et al., 1996). Artís et al. (2002) find a significant portion of the claims that were previously classified as legitimate contain omission errors, and thus are likely to be fraudulent. Further, Hausman et al. (1998) show that ignoring potential misclassification of a dependent variable can result in biased and inconsistent coefficient estimates when using standard parametric specifications.

Artís et al. (2002) present a logistic regression model accounting for misclassified claims and estimates it by the method of Hausman et al. (1998). Caudill et al. (2005) estimate this model by means of a multinomial logit model (MNL) and the EM-algortihm. They argue that identifying fraudulent claims is similar in nature to several other problems in real life including medical and epidemiological problems. They describe the methodology that can be used to produce parameter estimates with a dataset containing potentially misclassified dependent variables. Further, they estimate the proportion of fraudulent claims for car damage that are potentially erroneously classified as honest by an insurance company. The procedure is based on a transformation of the standard MNL likelihood function into a missing data formulation to which the expectation maximization (EM) algorithm can be applied (Dempster et al., 1977).

By assuming that the fraudsters that are caught have similar characteristics with the ones that are not, a latent variable model can be specified, where the group of those that were not caught in the initial audits is divided into two groups, the uncaught fraudsters and the truly honest claims. The model can then be estimated by the Expectation Maximization (EM) algorithm for missing data and be used to identify claims that probably are fraudulent, even if the audit did not catch them. This idea was introduced by Caudill et al. (2005), who fit the model to a dataset of Spanish car insurance fraud. Since this is real data, we do not know whether the EM-algorithm actually provides an improvement over other fraud detection methods, only that it is implementable.

Therefore, this paper investigates the methodology by means of a Monte Carlo study in order to evaluate its performance. We simulate data and vary the key relationships to evaluate the improvements that the EM-algorithm provides. We compare the parameter estimates obtained after running the EM-algorithm with the estimates obtained under a perfect information scenario. Additionally we compare the EM-parameters to a naive binomial logit model that does not take the misclassification into account. By doing so, we can see how much estimation accuracy we lose due to not having full information, and the improvement in performance over the naive approach.

The particular models we perform our simulation study on are guided by the empirical results of Caudill et al. (2005). The insurance claims categorized as honest, even though they are actually fraudulent, constitute the misclassified (missing) data. The data used by Caudill et al. (2005) is taken from Artís et al. (2002) and we use the standard deviations and coefficients from the paper by Artís et al. (2002) to simulate our data. In our simulated data we, of course, have full knowledge of whether a claim that is observed as being honest is really honest, or whether it is actually fraudulent.

The EM algorithm consists of two steps. In the Expectation (E) step, unobserved indicator variables associated with truly honest and honest-fraudulent claims are replaced with their conditional expectations, given the data and values of the unknown parameters. These conditional expectations or probabilities can be readily computed, given the structure of the logit model. In the Maximization (M) step, the log-likelihood function is maximized, new parameter values are obtained and then the E and M steps are repeated until the likelihood function is maximized. When the parameters are estimated, we can obtain the final estimates of the probabilities of whether a claim is fraudulent or not.

The EM-algorithm avoids the problem that the binomial logit model incurs, where all claims are assumed to be correctly classified; hence, the EM-algorithm avoids using misclassified observations for calculating probabilities, which is the cause of incorrect probabilities. Our results are aimed at revising claims initially classified as honest by reopening investigations and examining claims more closely, but might also improve prediction models. Further, this allows us to identify weaknesses in the initial classification system.

The paper is structured in the following way. Section 2 gives a literature review on theoretical and empirical studies of the detection of fraudulent claims. Section 3 presents the model by Caudill et al. (2005) and how to estimate it by means of the EM-algorithm. In Section 4 the performance of the EM-estimator is evaluated against two benchmark estimators by means of a Monte Carlo Study. A conclusion closes the paper.

2 The model

2.1 The multinomial model with missing information

The model by Caudill et al. (2005) is based on a multinomial distribution[1] with three categories. The first category, the honest honest (HH), are claims not caught in the audit process that are indeed not fraudulent. The second category, the honest fraudulent (HF), are fraudulent claims not caught in the audit process. The last category consists of fraudulent claims (F) caught in the audit process.

If all three categories were observed, it would simply be a trinomial logit model. Of course, this is not the case and the model was developed in order to allow for, and estimate the probability of undetected fraudulent claims. In order to do so, a similarity between the detected and undetected fraudulent claims will be assumed and reflected in a parameter restriction that will be imposed in the model. By denoting the number of HH as Y_1, the number of HF as Y_2 and the number of F as Y_3, we assume that

$$(Y_1, Y_2, Y_3) \sim \mathrm{MN}(1, (p_1, p_2, p_3)), \tag{1}$$

implying that

$$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = p_1^{y_1} \cdot p_2^{y_2} \cdot p_3^{y_3}. \tag{2}$$

The probability for an individual to belong to group 1, 2 or 3, respectively, is assumed to be given by the following equations

$$p_1 = P(Y_1 = 1, Y_2 = 0, Y_3 = 0) = \frac{1}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_3 x}}, \tag{3}$$

$$p_2 = P(Y_1 = 0, Y_2 = 1, Y_3 = 0) = \frac{e^{\alpha_2 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_3 x}}, \tag{4}$$

$$p_3 = P(Y_1 = 0, Y_2 = 0, Y_3 = 1) = \frac{e^{\alpha_3 + \beta_3 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_3 x}}. \tag{5}$$

In order to identify the parameters Caudill et al. (2005), assume that $\beta_2 = \beta_3$, i.e. that the probability of an individual to be fraudulent is affected by the explanatory variables in the same way, independent on whether the fraud has been detected or not. A difference in the probability for an individual to be in classes 2 or 3 is still allowed since α_2 and α_3 are free parameters.

[1] We denote this distribution $\mathrm{MN}(n, \boldsymbol{p})$, where n is the number of trials and $\boldsymbol{p}$ is a $K-$vector containing the probabilities for each of K categories.

2.2 Estimation of the model using EM algorithm

The α's and the β's are parameters to be estimated, and x is a vector of exogenous variables. We can now write the log-likelihood function

$$\ln L(\alpha_2, \alpha_3, \beta_2) = \sum_{i=1}^{n} (Y_{1i} \ln p_1 + Y_{2i} \ln p_2 + Y_{3i} \ln p_3), \tag{6}$$

where i $(i = 1, ..., n)$ represents an individual i, and n is the sample size. The ML-estimator is obtained by maximizing the log-likelihood with respect to the parameters α_2, α_3 and β_2.

We have only observed a binomial variable $(Z_2, Z_3) = (Y_1 + Y_2, Y_3)$, but model it as a trinomial (Y_1, Y_2, Y_3).

Since we do not observe all three categories we cannot compute $\ln L(\alpha_2, \alpha_3, \beta_2)$. We therefore use the suggested by Caudill et al. (2005), which is based on Expectation Maximization (EM) algorithm. Briefly described,

1. *Expectation (E) step*
   we compute the expectation of $\ln L(\alpha_2, \alpha_3, \beta_2)$ conditional on the observed data.

   $$Q(\alpha_2, \alpha_3, \beta_2) = E(\ln L(\alpha_2, \alpha_3, \beta_2)|\mathbf{Y}), \tag{7}$$

   where $\mathbf{Y}$ is an $n \times 3$-matrix where element (i, j) is equal to 1 if observation nr i belongs to category j and zero otherwise. Conditioning on the x-observations are also done but is not expressed explicitly in the formulas here. The conditional expectation of each term in (6) is computed by observing that only one of Z_2 and Z_1 is equal to one; the other is zero. We need

   $$\begin{aligned} Y_{2i}^* := E(Y_{2i}|Z_{2i} = 1) = P(Y_{2i} = 1|Z_{2i} = 1) &= \frac{p_2}{p_2 + p_3} \\ &= \frac{e^{\alpha_2 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x}} \end{aligned} \tag{8}$$

   and similarly

   $$Y_{1i}^* = \frac{1}{1 + e^{\alpha_2 + \beta_2 x}} \tag{9}$$

2. *Maximization (M) step*
   The log-likelihood function $\ln L(\alpha_1, \alpha_2, \beta_2)$ is maximized, where Y_{1i} and Y_{2i} are substituted by Y_{1i}^* and Y_{2i}^*. New α and β estimates are found, these are plugged in the step 1 and new Y_i^* estimates are found. Log-likelihood function is maximized again, based on these new values, and the whole process is iterated until the log-likelihood function is at its maximum.

In R, this can be implemented by the following algorithm:

1. Select starting values for $\alpha_2^{(1)}, \alpha_3^{(1)}, \beta_2^{(1)}$.

2. Compute $Y_{1i}^{*(1)}, Y_{2i}^{*(1)}$, based on these parameter values.

3. Maximize log-likelihood function where Y_{1i} and Y_{2i} are substituted with $Y_{1i}^{*(1)}$ and $Y_{2i}^{*(1)}$. After maximization, new parameter values $\alpha_2^{(2)}, \alpha_3^{(2)}, \beta_2^{(2)}$ are obtained.

4. Steps 2 to 3 are repeated until there is a convergence to the maximum likelihood estimator.

As starting values, we use the values obtained from estimating a binomial logit model, assuming no misclassification has been done.

3 Monte Carlo Study

In order to investigate how well the method manages to estimate the parameters of the model when some observations are incorrectly classified as honest, we perform a simulation study. The results can be seen as a best case scenario since we assume that we know the data generating process (DGP) except for the parameter values, which have to be estimated. In reality, the explanatory variables, and the functional form for the probabilities, will of course, not be specified exactly in accordance with the DGP.

Though simplified, in order to study realistic situations we have chosen the parameter values of the models guided by the study of Caudill et al. (2005). For the sake of a clear exhibition, we here recapitulate the model. Each individual claim is represented by a trinomial variable (Y_1, Y_2, Y_3) with zeros in two of the three entries and 1 in the class where the claim belongs. The probabilities are given by

$$p_1 = P(Y_1 = 1, Y_2 = 0, Y_3 = 0) = \frac{1}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_2 x}}, \tag{10}$$

$$p_2 = P(Y_1 = 0, Y_2 = 1, Y_3 = 0) = \frac{e^{\alpha_2 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_2 x}}, \tag{11}$$

$$p_3 = P(Y_1 = 0, Y_2 = 0, Y_3 = 1) = \frac{e^{\alpha_3 + \beta_2 x}}{1 + e^{\alpha_2 + \beta_2 x} + e^{\alpha_3 + \beta_2 x}}.^{2} \tag{12}$$

In the simulation study, performed in R (R Core Team, 2020)[3], we compare the results of the EM-estimators with two benchmarks. The first is to simply ignore that there is misclassification, i.e. to use a binomial logit model to estimate β_2; this estimator is denoted $\hat{\beta}_2^B$. This would be possible to do in practice. The second benchmark is to pretend that we actually observed all three categories and estimate β_2 by means of a trinomial logit model; this estimator is denoted $\hat{\beta}_2^T$. This is, obviously, not possible to do when the observations are

[2] Note that the restriction $\beta_2 = \beta_3$ is explicitly imposed in the model.

[3] The code can be found at `https://github.com/andreasolden/em_algorithm_missing_data`

only marked as ”caught” or ”not caught”. It is, however, a good comparison since we, with the EM-estimator, are trying to fit exactly the same model, but with a reduced form of the data. For all models and parameter combinations, 1000 replications have been performed to compute the Monte Carlo means and standard deviations. The sample size for each replication was 1000.

To compute the standard error, the estimated standard deviation, of the EM-estimator, the non-parametric bootstrap with 200 replications is used. 200 replications were chosen according to Tibshirani and Efron (1993), who show that running more than 200 replications provides very limited improvements in bootstrapped standard errors. An experiment with 1000 bootstrap replications for the standard errors was also conducted and did not indicate a noticeable improvement. For the trinomial and binomial logit model the standard errors are computed using the asymptotic distribution of the ML-estimators. Since there are no known closed form expressions for the standard deviations of the estimators, the performance of the standard errors is investigated by comparing their Monte Carlo mean with the Monte Carlo standard deviation of the estimator. In the tables, we call the latter the ”True SD”. Strictly speaking, this is of course only correct if the number of replications is infinitely large. For some replicates, the EM-algorithm did not converge after 100 iterations. However, we found no evidence that these estimates were systematically different from the ones that did. The tables presented below looked very similar when those replications were removed.

3.1 One explanatory variable

In this section we study the case with only one explanatory variable, x_1. The parameters of interest to estimate are therefore $(\alpha_2, \alpha_3, \beta_2)$. Guided by the empirical study in Caudill et al. (2005) we start by setting $(\alpha_2, \alpha_3, \beta_2) = (-1.8, -1.5, -0.02)$. The variable x_1 is thought of as the variable AGE in Caudill et al. (2005). The standard deviation of AGE is 12.3 and that is what we use as our starting case after which we also vary this quantity.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.670	0.095	0.084
$\hat{\alpha}_2^{T}$	-1.800	0.100	0.099
$\hat{\alpha}_2^{B}$	-1.664	0.087	0.087
$\alpha_3 =$ -1.5			
$\hat{\alpha}_3^{EM}$	-1.486	0.101	0.090
$\hat{\alpha}_3^{T}$	-1.505	0.087	0.088
$\beta_2 =$ -0.02			
$\hat{\beta}_2^{EM}$	-0.021	0.009	0.009
$\hat{\beta}_2^{T}$	-0.020	0.006	0.006
$\hat{\beta}_2^{B}$	-0.017	0.007	0.007

Table 1: Different estimators of α_2, α_3 and β_2 when the true values are $\alpha_2 = -1.8$, $\alpha_3 = -1.5$ and $\beta_2 = -0.02$ and $sd(x_1) = 12.3$. Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

From Table 1, we see that, with the exception of α_2, the EM-estimator performs almost as well as if all three categories would have been observed when $sd(x_1) = 12.3$. On the other hand, the mistake of ignoring misclassification, which is done by $\hat{\beta}_2^B$ is not that consequential. The bootstrapped standard errors are, on average, slightly underestimating the standard deviation of $\hat{\alpha}_2^{EM}$ and $\hat{\alpha}_3^{EM}$.

The benefit of the EM-estimator can, however, be seen in Table 2, where the standard deviation of the explanatory variable is increased with a factor of 10. The Monte Carlo mean of $\hat{\beta}_2^B$ is then -0.011 compared to the true value -0.020. The EM-estimator, $\hat{\beta}_2^{EM}$, has a Monte Carlo mean of -0.021. The estimation uncertainty is also, as expected, much smaller for the case with a large standard deviation in the explanatory variable. The estimation uncertainty, manifested in the Monte Carlo standard deviations (True SD), is of course larger than if we had observed all three categories.

As can be seen by comparing Table 1 in Table 2, and in all the remaining simulation experiments in the paper, $\hat{\alpha}_2^{EM}$ is less biased when the variance of the explanatory variable(s) is larger.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.824	0.360	0.353
$\hat{\alpha}_2^{T}$	-1.804	0.122	0.116
$\hat{\alpha}_2^{B}$	-1.700	0.095	0.104
$\alpha_3 =$ -1.5			
$\hat{\alpha}_3^{EM}$	-1.500	0.173	0.168
$\hat{\alpha}_3^{T}$	-1.503	0.108	0.108
$\beta_2 =$ -0.02			
$\hat{\beta}_2^{EM}$	-0.021	0.003	0.003
$\hat{\beta}_2^{T}$	-0.020	0.001	0.001
$\hat{\beta}_2^{B}$	-0.011	0.001	0.001

Table 2: Different estimators of α_2, α_3 and β_2 when the true values are $\alpha_2 = -1.8$, $\alpha_3 = -1.5$ and $\beta_2 = -0.02$ and $sd(x_1) = 123$. Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

3.2 Two explanatory variables

In order to investigate how the addition of more explanatory variables affects the results we now let $\mathbf{x} = (x_1, x_2)'$ and $\beta_2 = (\beta_{21}, \beta_{22})'$ be 2-dimensional vectors and the terms $\beta_2 x$ in equations (10)-(12) replaced by $\mathbf{x}'\beta_2$. The choice of parameter values studied is, again, guided by Caudill et al. (2005). x_1 is again thought of as representing the variable *AGE* and the x_2 variable *RECORDS*.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.680	0.214	0.256
$\hat{\alpha}_2^{T}$	-1.806	0.105	0.100
$\hat{\alpha}_2^{B}$	-1.671	0.088	0.089
$\alpha_3 =$ -1.5			
$\hat{\alpha}_3^{EM}$	-1.483	0.107	0.110
$\hat{\alpha}_3^{T}$	-1.504	0.088	0.089
$\beta_2 =$ -0.02			
$\hat{\beta}_2^{EM}$	-0.022	0.009	0.010
$\hat{\beta}_2^{T}$	-0.021	0.006	0.006
$\hat{\beta}_2^{B}$	-0.017	0.007	0.007
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.207	0.065	0.071
$\hat{\beta}_3^{B}$	0.202	0.041	0.041
$\hat{\beta}_3^{B}$	0.168	0.050	0.049

Table 3: Different estimators of α_2, α_3, β_2 and β_3 when the true values are $\alpha_2 = -1.8$, $\alpha_3 = -1.5$, $\beta_2 = -0.02$ and $\beta_3 = 0.2$, $sd(x_1) = 12.3$, $sd(x_2) = 1.8$ and $\mathrm{Corr}(x_1, x_2) = 0$. Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

As Table 3 shows, both the estimators and the standard errors are close to their respective true values. In the binomial model, the β-coefficients are biased towards zero due to the misspecification. Also for this case we investigate a situation with more variation in the explanatory variables. In Table 4, the standard deviations of x_1 and x_2 are both multiplied by 10.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.843	0.276	0.288
$\hat{\alpha}_2^{B}$	-1.808	0.140	0.136
$\hat{\alpha}_2^{B}$	-1.630	0.091	0.105
$\alpha_3 =$ -1.5			
$\hat{\alpha}_3^{EM}$	-1.524	0.193	0.205
$\hat{\alpha}_3^{T}$	-1.507	0.130	0.131
$\beta_2 =$ -0.02			
$\hat{\beta}_2^{EM}$	-0.020	0.003	0.004
$\hat{\beta}_2^{B}$	-0.020	0.002	0.002
$\hat{\beta}_2^{B}$	-0.007	0.001	0.001
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.205	0.031	0.035
$\hat{\beta}_3^{B}$	0.202	0.014	0.014
$\hat{\beta}_3^{B}$	0.072	0.006	0.006

Table 4: Different estimators of α_2, α_3, β_2 and β_3 when the true values are $\alpha_2 = -1.8$, $\alpha_3 = -1.5$, $\beta_2 = -0.02$ and $\beta_3 = 0.2$, $sd(x_1) = 123$, $sd(x_2) = 18$ and $\text{Corr}(x_1, x_2) = 0$. Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

We now go back to the case with the original standard deviations of x_1 and x_2, 12.3 and 1.8, respectively, but impose a correlation of 0.5 between them. The result is presented in Table 5. The difference with Table 3 is surprisingly small an we therefore investigated this by increasing the correlation to 0.9. This is presented in Table 6, which shows that the standard deviation of the estimators for β_2 and β_3 is larger.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.670	0.115	0.112
$\hat{\alpha}_2^{T}$	-1.804	0.104	0.100
$\hat{\alpha}_2^{B}$	-1.664	0.087	0.088
$\alpha_3 =$ -1.5			
$\hat{\alpha}_3^{EM}$	-1.483	0.101	0.096
$\hat{\alpha}_3^{T}$	-1.503	0.087	0.088
$\beta_2 =$ -0.02			
$\hat{\beta}_2^{EM}1$	-0.021	0.009	0.011
$\hat{\beta}_2^{T}$	-0.020	0.006	0.007
$\hat{\beta}_2^{B}$	-0.017	0.008	0.008
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.207	0.071	0.075
$\hat{\beta}_3^{T}$	0.203	0.046	0.047
$\hat{\beta}_3^{B}$	0.170	0.056	0.056

Table 5: Different estimators of α_2, α_3, β_2 and β_3 when the true values are $\alpha_2 = -1.8$, $\alpha_3 = -1.5$, $\beta_2 = -0.02$ and $\beta_3 = 0.2$, $sd(x_1) = 12.3$, $sd(x_2) = 1.8$ and $\mathrm{Corr}(x_1, x_2) = 0.5$. Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

Parameter	MC mean	True SD	Estimated SD
$\alpha_2 = -1.8$			
$\hat{\alpha}_2^{EM}$	-1.667	0.085	0.064
$\hat{\alpha}_2^{T}$	-1.803	0.102	0.099
$\hat{\alpha}_2^{B}$	-1.662	0.084	0.087
α_3 = -1.5			
$\hat{\alpha}_3^{EM}$	-1.485	0.097	0.089
$\hat{\alpha}_3^{T}$	-1.504	0.084	0.088
β_2 = -0.02			
$\hat{\beta}_2^{EM}$	-0.021	0.020	0.020
$\hat{\beta}_2^{T}$	-0.021	0.014	0.013
$\hat{\beta}_2^{B}$	-0.018	0.016	0.016
$\beta_3 = 0.2$			
$\hat{\beta}_3^{EM}$	0.206	0.134	0.135
$\hat{\beta}_3^{T}$	0.203	0.094	0.091
$\hat{\beta}_3^{B}$	0.173	0.113	0.111

Table 6: Different estimators of α_2, α_3, β_2 and β_3 when the true values are $\alpha_2 = -1.8$, $\alpha_3 = -1.5$, $\beta_2 = -0.02$ and $\beta_3 = 0.2$, $sd(x_1) = 12.3$, $sd(x_2) = 1.8$ and $\mathrm{Corr}(x_1, x_2) = 0.9$. Monte Carlo means and standard deviations of the estimators and the means of the standard errors.

To summarize the simulation study, for the most part, the EM-estimator estimates the parameters of the investigated data generating processes (DGP) well even though some observations are misclassified; with the exception of the EM-estimator of α_2, there are no indications that the parameter estimators are biased. In one experiment, identical to Table 1 but for the value of α_2 which was -3.0 instead of -1.8, the EM-estimator seem to systematically converge to value close to α_3 (-1.5) with the consequence that $\hat{\alpha}_2^{EM}$ was severely biased towards zero. This might be due to convergence to a local optimum, an hypothesis strengthened by a convergence close to the true value (-3.0), when the true values were used as starting values.

Bootstrapped standard errors also work well for the β-values in the investigated DPGs. However, the standard errors for the α-estimators underestimate the true standard deviations of the estimators when the variances of the explanatory variables are small. When the variances are large, the standard errors seem to overestimate the standard deviations of the estimators.

4 Conclusions

In this paper we investigate, by means of a Monte Carlo study, how well the EM-estimator of the model by Caudill et al. (2005) performs. We study different levels of variation in the explanatory variables in order to evaluate what is required to estimate the parameters well. In order to investigate realistic cases, we have chosen parameter values of the models guided by the study of Caudill et al. (2005) but simplified so that we study models with one or two explanatory variables only. In addition to investigating the point estimators of the parameter values we also study bootstrapped standard errors.

For the investigated models and parameter values, most point estimators work well, on average. The exception to this is the EM-estimator of α_2, which determines the difference between the correctly and incorrectly classified fraudulent observations. The estimator worked well when the variance of the explanatory variables was large. Overall, the bootstrapped standard errors also perform adequately as estimators of the standard deviations of the estimators. There is one exception also to this, though. When the variation in the explanatory variables is small, the standard deviation for $\hat{\alpha}_2^{EM}$ is underestimated and when the variance is large, it is overestimated.

We compare the estimators with two benchmarks. The first requires more data, namely that all three categories are observed. This trinomial model serves as an upper limit for how well the EM-estimator, which uses less information, could perform. The second benchmark is a binomial logit model where the fact that some observations are misclassified is ignored. The trinomial model, combined with observations of all three categories, as expected, captures the parameter values more precisely than the EM-estimator. The only interest in the results for the estimator of the binomial model is to show the effect of ignoring a part of the model (analogous to omitted variable bias in a linear regression). With small variance in the explanatory variables the bias is surprisingly small for the binomial estimator (for the cases when they can be seen as estimators of parameters in the trinomial model). This bias is exacerbated when the variance is increased.

References

J. Ai, P. L. Brockett, L. L. Golden, and Montserrat Guillén. A robust unsupervised method for fraud rate estimation. *The Journal of Risk and Insurance*, 80(1):121–143, 2013.

Jing Ai, Linda L Golden, and Patrick L Brockett. Assessing consumer fraud risk in insurance claims: An unsupervised learning technique using discrete and continuous predictor variables. *North American Actuarial Journal*, 13 (4):438–458, October 2009.

Manuel Artís, Mercedes Ayuso, and Montserrat Guillén. Modelling different types of automobile insurance fraud behaviour in the spanish market. *Insurance: Mathematics and Economics*, 24:67–81, 1999.

Manuel Artís, Mercedes Ayuso, and Montserrat Guillén. Detection of automobile insurance fraud with discrete choice models and misclassified claims. *Journal of Risk and Insurance*, 69(3):325–340, 2002.

P. L. Brockett, R. A. Derrig, L. L. Golden, A. Levine, and M. Alpert. Fraud classification using principal component analysis of ridits. *The Journal of Risk and Insurance*, 69(3):341–371, 2002.

Steven B Caudill, Mercedes Ayuso, and Montserrat Guillén. Fraud detection using a multinomial logit model with missing information. *Journal of Risk and Insurance*, 72(4):539–550, 2005.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

Richard A. Derrig and Krzysztof M. Ostaszewski. Fuzzy techniques of pattern recognition in risk and claim classification. *The Journal of Risk and Insurance*, 62(3):447–482, 1995.

J. A. Hausman, J. Abrevaya, and F. M. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87:239–269, 1998.

John A Major and Dan R Riedinger. Efd: A hybrid knowledge/statistical-based system for the detection of fraud. *Journal of Risk and Insurance*, 69 (3):309–324, September 2002.

Alan D Olinsky, Paul M Mangiameli, and Shaw K Chen. Statistical support of forensic auditing. *Interfaces*, 26(6):95–104, November 1996.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57:1–436, 1993.

S. Viaene, R. A. Derrig, B. Baesens, and G. Dedene. A comparison of state-of-the-art classification techniques for expert automobile insurance claim fraud detection. *The Journal of Risk and Insurance*, 69(3):373–421, 2002.

NHH

NORGES HANDELSHØYSKOLE
Norwegian School of Economics

Helleveien 30
NO-5045 Bergen
Norway

T +47 55 95 90 00
E nhh.postmottak@nhh.no
W www.nhh.no