Ordered Logit Analysis

For Selectively Sampled Data

By

Dennis FOK[1],
Philip Hans FRANSES[2] and
Mars CRAMER[3]

Econometric Institute Report 9933/A

Abstract

When customers are classified into ordered categories, which are defined from the outset, it may happen that the majority belongs to a single category. If a market researcher is interested in the correlation between the classification and individual characteristics, the natural question is whether one needs to collect data for all customers in that particular category. We address this question for the ordered logit model. We show that there is no need to consider all those customers. All that is required is a simple modification of the log-likelihood, which is based on Bayes' rule. We illustrate our proposed method on simulated data and on data concerning risk profiles of customers of an investment bank.

Key words: Ordered logit model, selective sampling, Bayes' rule

This version: August 10, 1999

1. Introduction

For marketing purposes it is often of interest to classify customers into various segments. For example, an investment firm, with access to a large database with

[1] Quantitative Research ROBECO Group and Erasmus Research Institute of Management. We thank Rabobank Nederland for providing us with the data.
[2] Corresponding author: Econometric Institute and Department of Marketing and Organisation, Office H11-15, Erasmus University Rotterdam, P.O.Box 1738, 3000 DR Rotterdam, fax: +3110 4089162, email: franses@few.eur.nl
[3] Tinbergen Institute Amsterdam

information on characteristics of its customers and their past investment behavior, may want to classify its customers according to risk profiles. Usually, these risk profiles are defined from the outset as discrete categories. Assuming there are *m* such categories, category 1 can contain the most risk-averse customers while categories 2 to *m* contain increasingly less risk-averse customers. The investment firm may now be interested in examining possible correlations between the characteristics of a customer and his or her classification into one of the risk profiles. As the variable to be explained is an ordered and discrete variable, one usually has to rely on an ordered regression model to summarize the correlations.

The database of an investment firm can be very large as it contains a host of information on oftentimes all customers. On the other hand, it may well be that only a few customers fall into one of the abovementioned categories. For example, supposing that there are N customers and $m = 3$ categories with N_1, N_2 and N_3 customers, N_1 and N_3 may be substantially outnumbered by N_2. A natural question is then whether one needs to include all N_2 individuals in the analysis of the ordered logit model. Indeed, if only a fraction of N_2 will suffice, one would save much time and effort, as not all data have to be collected, checked for errors, and stored. In this paper, we address this question for the ordered logit model. We will show that there is indeed no need to collect information on all N_2 customers, and that only a fraction will do. A modification of the likelihood function will give similar inference for both cases, that is, both estimates refer to the same parameters. In practice they do not differ substantially.

The outline of this paper is as follows. In Section 2, we briefly discuss some essentials of the ordered logit model. In Section 3, we put forward the modification of the log-likelihood, which allows for selective sampling from a large number of individuals who all would be classified into the same category. In Section 4, we evaluate our modification in a limited simulation experiment. In Section 5, we apply our method to risk profiles data from a large Dutch investment bank. In Section 6, we conclude with some remarks.

2. The ordered logit model

Consider the following dependent variable

$$y_{ij} = \begin{cases} 1 & \text{if customer } i \text{ belongs to category } j \\ 0 & \text{otherwise} \end{cases} \qquad i = 1, \cdots N,\ j = 1, \cdots m \tag{1}$$

where j can be thought of as a risk profile class, such that j= 1 corresponds with highly risk averse customers. Assume that there is a latent variable y^*, which can be modeled as

$$y_i^* = b'x_i + e_i \quad e_i \sim Logistic(q = 1) \tag{2}$$

that is, y^* can be explained by k explanatory variables contained in x (where x does not contain a column of ones for identification purposes). The logistic distribution with mean 0 has the following pdf

$$f(x) = \frac{1}{q} \frac{\exp(x/q)}{(1+\exp(x/q))^2} \tag{3}$$

The individuals are classified into the m categories by the following rule:

$$
\begin{aligned}
&y_{i,1} = 1 \qquad \text{if} \qquad y_i^* \le a_1 \\
&y_{i,j} = 1 \qquad \text{if } a_{j-1} < y_i^* \le a_j \text{ for } j = 2,..,m-1 \\
&y_{i,m} = 1 \qquad \text{if } a_{m-1} < y_i^*
\end{aligned} \tag{4}
$$

The thresholds a_i must satisfy $a_1 < a_2 < \ldots < a_{m-1}$. When we introduce $a_0 = -\infty$ and $a_m = +\infty$, customer i belongs to category j if $a_{j-1} < y_i^* \le a_j$, $j = 1,\ldots, m$. Combining (1), (2) and (3), we obtain that

$$
\begin{aligned}
&P(\text{customer } i \text{ belongs to category } j) = P(y_{ij} = 1) \\
&\quad = P(a_{j-1} < y_i^* \le a_j) = P(a_{j-1} < b'x_i + e_i \le a_j) \\
&\quad = P(a_{j-1} - b'x_i < e_i \le a_j - b'x_i) \\
&\quad = F(a_j - b'x_i) - F(a_{j-1} - b'x_i)
\end{aligned} \tag{5}
$$

where F denotes the cumulative density function of the logistic distribution. This model in (5) is called the ordered logit model, see for example McKelvey and Zavoina (1975) and McCullagh (1980) for some early applications.

The parameters of the model can be estimated using the maximum likelihood technique. The likelihood function follows directly from equation (5), that is,

$$
\begin{aligned}
L(a, b) = &\prod_{\substack{i,j \\ \text{client } i \text{ in class } j}} P(y_{ij} = 1) = \prod_{i=1}^{n}\prod_{j=1}^{m} P(y_{ij} = 1)^{y_{ij}} = \\
&= \prod_{i=1}^{n}\prod_{j=1}^{m} \left[F(a_j - b'x_i) - F(a_{j-1} - b'x_i)\right]^{y_{ij}}
\end{aligned} \tag{6}
$$

The parameters are estimated by maximizing the log-likelihood, given by

$$
\ln L = \sum_i \sum_j y_{ij} \ln\left[F(a_j - b'x_i) - F(a_{j-1} - b'x_i)\right] \tag{7}
$$

We use the Conjugate-Gradient numerical optimization algorithm to maximize the log-likelihood. This technique is available in a Matlab toolbox by the Numerical Algorithms Group Ltd. The method performs successive line minimizations along conjugate directions. The algorithm uses the derivatives of the log likelihood. To save notation, we write

$$
\begin{aligned}
F_{ij} &= F(a_j - b'x_i) \\
f_{ij} &= f(a_j - b'x_i)
\end{aligned}
$$

The derivatives are now given by

$$
\begin{aligned}
&\frac{\partial \ln L}{\partial b} = \sum_i \sum_j y_{ij} \frac{f_{i,j-1} - f_{ij}}{F_{ij} - F_{i,j-1}} x_i \\
&\frac{\partial \ln L}{\partial a_k} = \sum_i y_{ik} \frac{f_{ik}}{F_{ik} - F_{i,k-1}} - y_{i,k+1} \frac{f_{ik}}{F_{i,k+1} - F_{i,k}} \qquad k = 1,\ldots, m-1
\end{aligned} \tag{8}
$$

Unrestricted optimization of the log-likelihood does not guarantee a feasible solution. In fact, the estimated thresholds should obey the restriction $a_1 < a_2 < \ldots < a_{m-1}$. To make sure this restriction is satisfied we can make two adjustments. One possibility is to introduce a penalty in the log-likelihood. When the restriction is not satisfied we add a large negative value to the likelihood. This adjustment makes sure that an infeasible parameter configuration cannot maximize the log-likelihood. The second possibility, which seems more convenient, amounts to using a parameter transformation. That is, we transform the parameters of the threshold in such a way that the restriction is always satisfied. Instead of maximizing over a we maximize the likelihood over m, where

$$\begin{aligned} a_1 &= m_1 \\ a_2 &= m_1 + m_2^2 = a_1 + m_2^2 \\ a_3 &= m_1 + m_2^2 + m_3^2 = a_2 + m_3^2, \ldots\ldots \end{aligned} \tag{9}$$

Note that this transformation implies that $a_1 \le a_2 \le \ldots \le a_{m-1}$. When we maximize over m instead of a, we need the derivatives of the log-likelihood with respect to m. The derivatives in (8) have to be replaced with

$$\frac{\partial \ln L}{\partial m_k} = \sum_i \frac{\partial \ln L}{\partial a_i}\frac{\partial a_i}{\partial m_k}$$

$$\frac{\partial a}{\partial m} = \begin{bmatrix} \frac{\partial a_1}{\partial m_1} & \frac{\partial a_1}{\partial m_2} & \cdots & \cdots \\ \frac{\partial a_2}{\partial m_1} & \frac{\partial a_2}{\partial m_2} & \cdots & \cdots \\ \vdots & \vdots & \ddots & \\ \vdots & \vdots & & \ddots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ 1 & 2m_2 & 0 & 0 & \cdots \\ 1 & 2m_2 & 2m_3 & 0 & \cdots \\ 1 & 2m_2 & 2m_3 & 2m_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{10}$$

The variance of the maximum likelihood estimators can be estimated by the inverse of the information matrix

$$Est.Var(q) = I(q)^{-1} = \left[E\left(-\frac{\partial^2 \ln L}{\partial q \partial q'}\right) \right]^{-1} \quad \text{, where } q = \begin{bmatrix} b \\ m \end{bmatrix} \text{ or } \begin{bmatrix} b \\ a \end{bmatrix} \tag{11}$$

Estimates of this variance can be obtained using the analytical second order derivatives, but we choose to calculate the derivatives numerically. The numerical second derivatives are obtained from the Taylor expansion of the first derivatives, that is,

$$\left.\frac{\partial^2 \ln L}{\partial q \partial q_i}\right|_{q=\hat{q}} \approx \frac{\left.\frac{\partial \ln L}{\partial q}\right|_{q=\hat{q}, q_i=\hat{q}_i+h} - \left.\frac{\partial \ln L}{\partial q}\right|_{q=\hat{q}, q_i=\hat{q}_i-h}}{2h} \tag{12}$$

The application of this formula will give a vector of derivatives. To obtain the full matrix of second derivatives, the above formula should be applied to all parameters. When all vectors of derivatives are stacked, an estimate of the second order derivatives (Hessian) is obtained. Although the likelihood is a continuous differentiable function, the estimated Hessian is not always symmetric. This is due to

round-off errors and the approximation of the derivatives. A symmetric estimate can be obtained by applying

$$H_{new} = \tfrac{1}{2}(H + H') \tag{13}$$

where H denotes the Hessian.

To obtain accurate estimates from the optimization algorithm it usually helps to standardize the regressors. The mean is subtracted from the regressors and then one divides by the standard deviation. To obtain the estimates of the original coefficients, the following transformations have to be made

$$\begin{aligned} b_k &\leftarrow \frac{b_k}{s_k} \\ a &\leftarrow a - \sum_k \frac{b_k m_k}{s_k} \end{aligned} \tag{14}$$

In the next section, we examine how we have to modify the log-likelihood to obtain proper estimates in case we reduce the sample.

3. Selective sampling and the ordered logit model

When a market researcher makes an endogenous selection of the available observations, the estimation method needs to be adjusted. This adjustment follows from an application of the theorem of Bayes, see also Cramer, Franses and Slagter (1999) for related results for a censored regression model. Recall that when we assume an ordered logit model, the true probabilities in the population for customer i and category j are

$$P_{ij} = P(y_{ij} = 1 \mid x_i) = F(a_j - b'x_i) - F(a_{j-1} - b'x_i) \tag{15}$$

When the full sample is a random sample from the population with sampling fraction a, the probabilities that individual i is in the observed sample and is a member of class 1, 2, to m are then

$$aP_{i1}, \quad aP_{i2}, \quad \cdots, \quad aP_{im} \tag{16}$$

These probabilities do not sum to 1 because it is also possible that an individual is not present in the sample, which happens with probability (1-a). If however the number of observations in class j is reduced by g_j where the deleted observations are selected at random, these probabilities become

$$a\,g_1 P_{i1}, \qquad a\,g_2 P_{i2}, \qquad \cdots, \qquad a\,g_m P_{im} \tag{17}$$

Of course, when all observations are kept, then $g_j = 1$. To simplify notation, we collect the reduction factors in the vector Γ and the true population probabilities P_{ij} in the matrix P_i, that is

$$\Gamma = \begin{pmatrix} g_1 \\ \vdots \\ g_m \end{pmatrix} \qquad P_i = \begin{pmatrix} P_{i1} \\ \vdots \\ P_{im} \end{pmatrix}$$

The probability of observing $y_{ij} = 1$ in the reduced sample is now given by

$$\tilde{P}_{ij} = \frac{a\ g_j P_{ij}}{a\ g_1 P_{i1} + a\ g_2 P_{i2} + \cdots + a\ g_m P_{im}} = \frac{g_j P_{ij}}{\Gamma' P_i} \tag{18}$$

When we apply Bayes' theorem directly we obtain the same result, that is

$$\begin{aligned} \tilde{P}_{ij} &= P(y_{ij} = 1 \mid \text{customer } i \text{ is observed}) \\ &= \frac{P(\text{customer } i \text{ is observed} \mid y_{ij} = 1) P(y_{ij} = 1)}{P(\text{customer } i \text{ is observed})} \\ &= \frac{a\ g_j \cdot P_{ij}}{\sum_k P(\text{customer } i \text{ is observed} \mid y_{ik} = 1) P(y_{ik} = 1)} = \frac{a\ g_j P_{ij}}{\sum_k a\ g_k P_{ik}} = \frac{g_j P_{ij}}{\Gamma' P_i} \end{aligned} \tag{19}$$

With these adjusted probabilities, we can construct the new likelihood (and log-likelihood) function as follows:

$$\begin{aligned} L &= \prod_{\substack{i,j \\ \text{customer } i \text{ belongs to } j}} P(y_{ij} = 1 \mid \text{customer } i \text{ is observed}) = \prod_{i=1}^{n} \prod_{j=1}^{m} \tilde{P}_{ij}^{\ y_{ij}} \\ \ln L &= \sum_i \sum_j y_{ij} \ln\left(\frac{g_j}{\Gamma' P_i} P_{ij}\right) \end{aligned} \tag{20}$$

To optimize the likelihood we need the derivatives of the log-likelihood to the parameters, where it should be noted that the constants in Γ are known. Writing q for b or a_k, we have

$$\frac{\partial \ln L}{\partial q} = \sum_i \sum_j y_{ij} \left[\frac{\frac{\partial P_{ij}}{\partial q} \Gamma' P_i - \Gamma' \frac{\partial P_i}{\partial q} P_{ij}}{P_{ij} \Gamma' P_i} \right] \tag{21}$$

with

$$\begin{aligned} P_{ij} &= F(a_j - b' x_i) - F(a_{j-1} - b' x_i) = F_{ij} - F_{i,j-1} \\ \frac{\partial P_{ij}}{\partial b} &= -\left[f(a_j - b' x_i) - f(a_{j-1} - b' x_i) \right] x_i = (f_{i,j-1} - f_{i,j}) x_i \\ \frac{\partial P_{ij}}{\partial a_k} &= d_{j,k} f(a_j - b' x_i) - d_{j-1,k} f(a_{j-1} - b' x_i) = d_{j,k} f_{i,j} - d_{j-1,k} f_{i,j-1} \\ d_{l,k} &= \begin{cases} 1 & \text{if } l = k \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{22}$$

When $g_j=1$ for all j, no observations are deleted, and in that case we have

$$\Gamma' P_i = \sum_j P_{ij} = 1 \text{ and } \Gamma' \frac{\partial P_i}{\partial q} = 0, \tag{23}$$

This matches exactly with the derivative in the ordered logit model without selective sampling.

Maximizing the log-likelihood in (20) gives estimates of the relevant parameters. It is difficult to derive any exact results for the effect of reducing the samples. Therefore, we analyze our method for simulated data in the next section.

4. Simulation results

To evaluate the practical usefulness of our method which involves a correction of the likelihood function, we consider the following model to generate realizations of an ordered logit process:

$$y_i^* = 4x_{1i} + 2x_{2i} - 1x_{3i} + 3e_i \qquad e_i \sim Logistic$$

$$\begin{aligned} x_{1i} &\sim N(m=0, s=2) \\ x_{2i} &\sim N(m=0, s=4) \\ x_{3i} &\sim N(m=0, s=1) \end{aligned} \tag{24}$$

$$\text{category of customer } i = \begin{cases} \text{I} & y_i^* \le -15 \\ \text{II} & -15 < y_i^* \le 15 \\ \text{III} & 15 < y_i^* \end{cases}$$

Model (24) implies that approximately 77% of the individuals can be classified into category II, where categories I and III both contain approximately 11.5% of the observations. Hence, in a random sample the individuals in category II outnumber those in the other two groups. We use different factors to reduce the observations for group II to investigate the effect of this reduction on the precision of the parameter estimates. The estimates based on the full sample are compared to naïve estimates, which are calculated using the *standard* method based on the *reduced* sample, and to estimates obtained from the adjusted likelihood, also based on the *reduced* sample.

The full sample consists of 5000 observations and we generate 100 replications. Before we analyze all parameters, we present some typical results in Figure 1. Figure 1 depicts the average estimates of the second threshold parameter over 100 replications for each method and reduction factor. The dashed line gives the average estimate for the naïve method. The solid lines give the estimates based on the full sample and the estimates based on the adjusted method, which are of course based on the reduced sample. It is quite clear that the estimate of the second threshold is biased for the naïve method. In contrast, the adjusted method does provide an unbiased estimate. Notice that we would expect the line for the full sample to be a straight line. However, for every reduction factor we generate new data, and therefore the estimates

based on the full sample also differ. Naturally, when the number of replications or the number of observations is very large, these differences will become very small.

Table 1 gives the estimates of all parameters for all reduction factors. It is clear that the estimates of the thresholds obtained from the naïve method are highly biased, although the regressor coefficient estimates (slopes) are all estimated quite close to the true values. As expected, our adjusted method gives unbiased estimates, while also the slopes are closer to the true values than estimates by the naïve method.

However, we have to be careful in drawing general conclusions from one simulation experiment. Therefore, we repeat our simulation study for the case with thresholds equal to 5 and 10 and we reduce the number of observations in the first category (which contains roughly 67% of the observations). Table 2 shows that, when we reduce the observations in the first category, almost all coefficients are incorrectly estimated by the naïve method. The thresholds are again affected the most, with the first threshold of the naïve method often taking values smaller than 4 whereas the true value is 5. For this data generating process, the bias of the estimated slopes is larger than for the previous process. Clearly, the adjusted method outperforms the naïve method.

It is also of interest to investigate the effect of the reduction on the absolute errors of the estimates. For this purpose, we can use the mean squared error (MSE) to compare the results. The MSE measures the average squared deviation of the estimate from the true value, that is,

$$\text{Mean squared Error}(q) = MSE(q) = \frac{1}{\#\,replications} \sum_{replications} (\hat{q} - q)^2 \qquad (25)$$

Upon using this MSE we can calculate an empirical confidence interval around the estimates, with the width of this interval measuring the accuracy of the estimates. We assume that the center of the interval is the true parameter value. So the intervals are based on the assumption of unbiasedness of the estimates.

We conclude from Table 3 that the confidence intervals, for the method based on the adjusted log-likelihood, are slightly wider than the intervals based on analyzing the full sample. This is of course due to the fact that the adjusted method uses less data. The intervals for the slopes do not differ that much between the naïve and the adjusted method, but the intervals for the adjusted method are in general less wide. For the second model (see Table 4) almost the same conclusions can be drawn. The difference between the performance of the naïve and the adjusted method is even larger. The intervals for the estimates of the first two slope coefficients with the naïve method are often more than two times as large as the intervals based on the full sample. Using the adjusted method the intervals for all parameters are less than 1.5 times as large. Additionally, these last intervals compare favorably with the intervals one would obtain if the rule of the square root of the number of observations would apply.

Based on the outcomes of this limited simulation experiment it is difficult to find general rules for the most appropriate reduction factor. The adjusted method will give good estimates for a wide range of reduction factors. As in an ordered logit model all observations contain an equal amount of information, we might want to use the general rule of equally sized groups.

5. An application to risk profiles

In this section we illustrate the method based on the adjusted log-likelihood for real-life data. Our potential data set consists of 41582 customers, who can be classified as having a low, middle or high risk profile. From Table 5, one can see that most (that is, almost 98%) individuals are classified into the middle category and that only about 850 individuals are classified as having low or high risk profiles. For each of these customers we have information on 9 explanatory variables, where we should mention that these have not been used to determine the classification. These variables appear on the left–hand side of Table 6. Due to confidentiality reasons, we cannot provide further details on the variables, except that they can concern the current state of a customer (for example, number of type I funds) or the behaviour in the recent past (for example, number of type I transactions).

The estimation results for an ordered logit model, where we consider various reduction factors for the observations concerning the middle category, are reported in Table 6. From this table, we can conclude that considering 10% or 20% of the 40772 observations yields approximately the same parameter estimates (and not very different standard errors) as in case we analyze the model for all individuals. This is even more clear from the relative parameter values given in Table 7. Only for the variable "Number of type III transactions" (which concern transactions on a high risk type of financial product), we observe that substantial differences appear. In Table 8, we demonstrate that this is most likely due to aberrant observations in the middle category. A few individuals in this category have exceptionally large values for this variable.

.

6. Concluding remarks

We proposed a simple modification to the log-likelihood of an ordered logit model, which enables a market researcher to discard a large number of observations from a category containing substantially more observations than other categories do. Through Monte Carlo simulations and an analysis of real-life data, we showed that our method results in unbiased estimates and that the estimated standard deviations do not increase to a large extent. Our method is useful for practical purposes as one may save on collecting, checking and storing data.

Our empirical analysis highlighted that outliers can have a large effect on the final results. As expected, such observations become less influential in a very large data set, and their effect becomes more pronounced if the market researcher is unlucky enough to select these observations for the reduced sample. Hence, a further topic for research is to consider methods that can help to prevent such unfortunate selections.

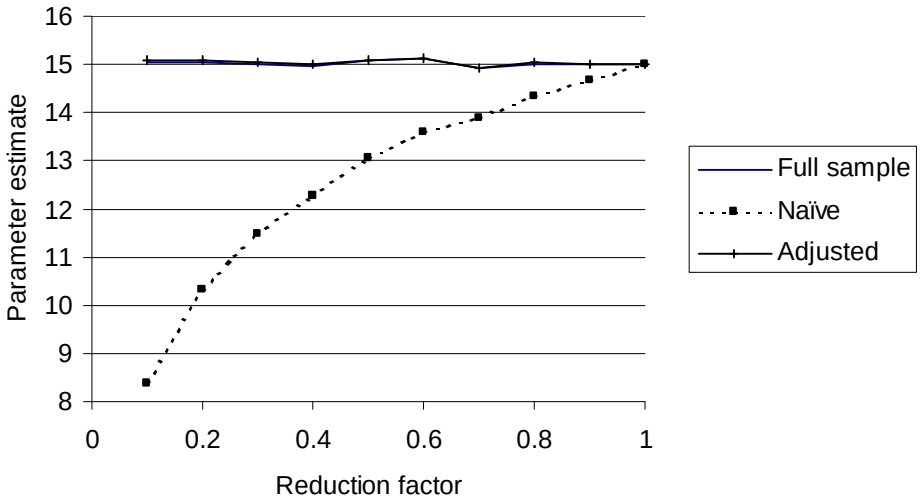

Figure 1: Estimates of α_2 for all three methods and for different reduction factors

	α_1 (-15)			α_2 (15)			β_1 (4)			β_2 (2)			β_3 (-1)		
Reduction	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted
0.1	-15.053	-8.391	-15.076	15.033	8.390	15.076	4.007	4.094	4.021	2.005	2.048	2.011	-1.010	-1.033	-1.014
0.2	-15.000	-10.296	-15.037	15.029	10.328	15.069	4.011	4.053	4.026	1.999	2.025	2.011	-1.020	-1.023	-1.016
0.3	-15.007	-11.460	-15.025	15.020	11.483	15.048	4.017	4.059	4.044	2.001	2.015	2.008	-1.013	-1.033	-1.030
0.4	-14.974	-12.290	-15.011	14.977	12.281	15.002	3.987	4.010	4.002	2.002	2.015	2.010	-1.029	-1.032	-1.029
0.5	-15.129	-13.097	-15.160	15.097	13.039	15.102	4.022	4.036	4.031	2.016	2.023	2.021	-1.014	-1.013	-1.012
0.6	-15.171	-13.655	-15.175	15.119	13.603	15.123	4.039	4.042	4.039	2.020	2.023	2.021	-0.999	-0.998	-0.997
0.7	-14.940	-13.877	-14.940	14.929	13.874	14.937	3.980	3.985	3.983	1.991	1.992	1.991	-0.995	-0.992	-0.992
0.8	-15.048	-14.378	-15.042	15.025	14.361	15.026	4.009	4.011	4.010	2.010	2.011	2.011	-1.010	-1.005	-1.005
0.9	-15.009	-14.703	-15.017	14.996	14.685	15.000	4.001	4.006	4.005	1.995	1.997	1.997	-0.980	-0.981	-0.981
1	-14.991	-14.991	-14.991	14.992	14.992	14.992	3.998	3.998	3.998	2.003	2.003	2.003	-1.013	-1.013	-1.013

Table 1: Average parameter estimates with the true parameter value in parentheses

	α_1 (5)			α_2 (10)			β_1 (4)			β_2 (2)			β_3 (-1)		
Reduction	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted
0.1	4.990	-1.936	5.013	9.995	7.190	10.006	3.993	3.436	4.001	2.004	1.719	2.002	-1.027	-0.858	-0.991
0.2	4.985	0.083	4.990	9.982	7.842	9.992	3.997	3.616	3.999	2.000	1.811	2.004	-1.019	-0.924	-1.020
0.3	5.015	1.318	5.016	10.051	8.346	10.043	4.010	3.728	4.008	2.009	1.865	2.004	-0.993	-0.920	-0.992
0.4	4.982	2.158	4.980	9.987	8.650	9.991	3.991	3.787	3.992	2.002	1.903	2.006	-0.995	-0.933	-0.983
0.5	5.023	2.883	5.025	10.065	9.039	10.080	4.032	3.885	4.038	2.018	1.949	2.026	-0.996	-0.967	-1.004
0.6	4.998	3.414	4.995	10.033	9.259	10.045	3.996	3.900	4.009	2.000	1.948	2.003	-0.982	-0.954	-0.981
0.7	5.038	3.939	5.045	10.037	9.483	10.047	4.007	3.935	4.010	2.010	1.975	2.013	-1.013	-0.994	-1.012
0.8	4.999	4.307	5.000	10.009	9.657	10.014	4.019	3.979	4.024	2.008	1.986	2.008	-0.998	-0.992	-1.003
0.9	4.991	4.661	4.989	9.992	9.820	9.992	3.997	3.977	3.998	2.000	1.990	2.001	-1.011	-1.005	-1.010
1	5.011	5.011	5.011	10.021	10.021	10.021	4.007	4.007	4.007	2.007	2.007	2.007	-1.001	-1.001	-1.001

Table 2: Parameter estimates when reducing first group

	α_1 (-15)			α_2 (15)			β_1 (4)			β_2 (2)			β_3 (-1)		
Reduction	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted
0.1	0.838	13.255	1.069	0.896	13.261	1.125	0.276	0.447	0.428	0.123	0.213	0.202	0.263	0.479	0.467
0.2	0.798	9.457	0.994	0.807	9.397	1.032	0.252	0.369	0.365	0.121	0.188	0.186	0.289	0.436	0.433
0.3	0.786	7.137	0.919	0.780	7.089	0.900	0.241	0.351	0.346	0.131	0.145	0.144	0.273	0.359	0.357
0.4	0.834	5.498	0.935	0.812	5.510	0.902	0.262	0.301	0.302	0.127	0.152	0.152	0.285	0.376	0.374
0.5	0.799	3.899	0.912	0.799	4.014	0.883	0.237	0.266	0.264	0.126	0.152	0.152	0.279	0.323	0.323
0.6	0.888	2.824	0.929	0.839	2.921	0.890	0.232	0.255	0.253	0.130	0.143	0.142	0.261	0.305	0.305
0.7	0.830	2.406	0.873	0.768	2.398	0.836	0.235	0.249	0.251	0.124	0.131	0.132	0.300	0.304	0.304
0.8	0.666	1.427	0.704	0.706	1.465	0.720	0.207	0.223	0.223	0.095	0.101	0.100	0.281	0.291	0.291
0.9	0.902	1.093	0.919	0.878	1.083	0.883	0.248	0.257	0.257	0.133	0.136	0.137	0.320	0.318	0.318
1	0.784	0.784	0.784	0.746	0.746	0.746	0.245	0.245	0.245	0.112	0.112	0.112	0.244	0.244	0.244

Table 3: Width of empirical confidence interval ($2\sqrt{MSE(q)}$), all three methods.
Results are based on first data generating process.

	α_1 (5)			α_2 (10)			β_1 (4)			β_2 (2)			β_3 (-1)		
Reduction	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted	Full	Naïve	Adjusted
0.1	0.363	13.881	0.523	0.475	5.657	0.588	0.195	1.158	0.273	0.099	0.576	0.142	0.265	0.415	0.339
0.2	0.325	9.844	0.458	0.516	4.368	0.610	0.212	0.815	0.288	0.105	0.399	0.139	0.245	0.347	0.339
0.3	0.368	7.376	0.433	0.505	3.357	0.528	0.214	0.588	0.233	0.110	0.296	0.128	0.241	0.333	0.312
0.4	0.341	5.698	0.409	0.442	2.743	0.458	0.188	0.473	0.212	0.108	0.225	0.117	0.230	0.290	0.275
0.5	0.355	4.250	0.385	0.548	2.013	0.581	0.247	0.348	0.278	0.118	0.162	0.138	0.246	0.288	0.289
0.6	0.328	3.191	0.345	0.532	1.588	0.549	0.218	0.302	0.231	0.098	0.148	0.108	0.295	0.314	0.312
0.7	0.371	2.151	0.375	0.518	1.170	0.541	0.218	0.271	0.242	0.114	0.131	0.126	0.242	0.274	0.280
0.8	0.338	1.427	0.344	0.457	0.841	0.478	0.204	0.214	0.217	0.104	0.115	0.114	0.262	0.256	0.258
0.9	0.370	0.775	0.378	0.474	0.605	0.483	0.222	0.231	0.228	0.094	0.099	0.098	0.260	0.274	0.276
1	0.371	0.371	0.371	0.509	0.509	0.509	0.213	0.213	0.213	0.111	0.111	0.111	0.228	0.228	0.228

Table 4: Width of empirical confidence interval.
Results are based on second data generating process.

	No. Obs	Percentage
Low	531	1.28%
Middle	40,772	97.93%
High	329	0.79%

Table 5: Number of observations and percentages in full sample

	Reduction factor									
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1
No. type I funds	0.072	0.066	0.057	0.059	0.060	0.062	0.064	0.063	0.062	0.063
	(3.068)	(2.730)	(2.434)	(2.689)	(2.712)	(2.773)	(2.915)	(2.867)	(2.840)	(2.868)
No. type I transactions	-0.043	-0.041	-0.034	-0.037	-0.042	-0.037	-0.037	-0.038	-0.038	-0.038
	(-3.264)	(-3.219)	(-2.964)	(-3.196)	(-3.554)	(-3.200)	(-3.274)	(-3.445)	(-3.369)	(-3.369)
Ind. type Ia funds	0.545	0.542	0.493	0.482	0.472	0.472	0.455	0.464	0.471	0.474
	(4.264)	(4.230)	(3.925)	(3.914)	(3.795)	(3.816)	(3.704)	(3.800)	(3.861)	(3.885)
No. type II funds	0.086	0.090	0.088	0.088	0.096	0.085	0.078	0.082	0.080	0.079
	(7.231)	(7.894)	(8.680)	(8.743)	(9.257)	(8.793)	(8.828)	(9.072)	(9.079)	(9.075)
Ind. type II funds	0.534	0.490	0.440	0.447	0.419	0.447	0.471	0.454	0.474	0.479
	(4.140)	(3.883)	(3.566)	(3.634)	(3.397)	(3.654)	(3.883)	(3.766)	(3.936)	(3.981)
No. type III transactions	0.186	0.130	0.132	0.101	0.066	0.101	0.029	0.065	0.028	0.029
	(9.219)	(8.938)	(10.262)	(9.167)	(6.541)	(9.949)	(5.588)	(6.368)	(6.134)	(6.246)
Turnover	0.002	0.003	0.002	0.002	0.002	0.002	0.003	0.002	0.003	0.003
	(1.906)	(2.932)	(2.675)	(2.627)	(3.287)	(3.145)	(4.150)	(3.703)	(4.261)	(4.374)
Wealth measure	0.023	0.020	0.021	0.021	0.022	0.021	0.019	0.023	0.021	0.021
	(2.182)	(2.324)	(2.505)	(2.757)	(2.926)	(2.765)	(2.532)	(2.966)	(2.762)	(2.788)
No. of special accounts	0.540	0.561	0.600	0.644	0.673	0.646	0.725	0.698	0.737	0.741
	(5.109)	(5.400)	(5.921)	(6.346)	(6.731)	(6.255)	(7.302)	(6.976)	(7.423)	(7.464)
α_1	-3.455	-3.488	-3.546	-3.548	-3.561	-3.560	-3.557	-3.557	-3.547	-3.544
α_2	6.199	6.123	6.055	6.024	6.004	6.010	5.961	5.985	5.981	5.982

Table 6: Parameters and *t*-values in parentheses for different reduction factors

	Reduction factor									
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1
No. type I funds	1.139	1.043	0.913	0.945	0.962	0.989	1.020	0.999	0.989	1
No. type I transactions	1.143	1.081	0.891	0.991	1.126	0.973	0.970	1.022	1.000	1
Ind. type Ia funds	1.150	1.145	1.041	1.018	0.997	0.996	0.959	0.980	0.995	1
No. type II funds	1.089	1.145	1.122	1.116	1.223	1.078	0.992	1.036	1.017	1
Ind. type II funds	1.114	1.023	0.919	0.933	0.875	0.932	0.983	0.948	0.989	1
No. type III transactions	6.480	4.541	4.580	3.512	2.288	3.507	1.012	2.261	0.981	1
Turnover	0.762	1.335	0.750	0.681	0.894	0.843	0.981	0.933	0.978	1
Wealth measure	1.122	0.943	1.032	1.010	1.071	0.998	0.905	1.109	0.996	1
No. of special accounts	0.729	0.758	0.810	0.870	0.908	0.872	0.979	0.942	0.995	1
α_1	0.975	0.984	1.001	1.001	1.005	1.005	1.004	1.004	1.001	1
α_2	1.036	1.024	1.012	1.007	1.004	1.005	0.996	1.000	1.000	1

Table 7: Parameters relative to estimates in full sample

Number of	Number of observations		
transactions	Low	Middle	High
> 25	2	31	14
> 50	0	12	5
> 75	0	7	5
> 100	0	6	3
> 400	0	1	1

Table 8: Customers with many type III transactions

References

Cramer, Mars, Philip Hans Franses and Erica Slagter (1999), Censored regression analysis in large samples with very many zero observations, Unpublished manuscript, Tinbergen Institute Amsterdam.

McCullagh, Peter (1980), Regression models for ordinal data (with discussion), Journal of the Royal Statistical Society B42, 109-142.

McKelvey, Richard D. and William Zavoina (1975), A statistical model for the analysis of ordinal level dependent variables, Journal of Mathematical Sociology 4, 103-120.