Censored Regression Analysis in Large Samples with Many Zero Observations

Mars Cramer, Philip Hans Franses and Erica Slagter*

November 8, 1999

ECONOMETRIC INSTITUTE RESEARCH REPORT 9939/A

Abstract

With the advent of improved data collection techniques, the applied econometrician can nowadays have access to very large data bases. Sometimes, however, these can have fairly low informational content. For example, a typical response rate in direct mailings is below 1%. Given the small fraction of respondents, one could be tempted to omit the larger part of the non-respondents from the analysis. If so, one should adapt the statistical analysis to this new situation. We put forward such an adaptation for the censored regression model. This model is often used in marketing research, for example, to analyze the amount of money spent on new products offered in a direct mailing campaign. We discuss how the likelihood function should be modified to obtain proper maximum likelihood [ML] estimates. Our empirical illustration concerns a data set of about 300000 observations. We show that our modified ML method yields the appropriate estimates, and that the loss of efficiency is not large.

Keywords: Logit model, censored regression, selective sampling

*Cramer is affiliated with the Tinbergen Institute, Amsterdam, Franses with the Department of Marketing and Organization and the Econometric Institute, both of the Erasmus University Rotterdam, and Slagter with the Robeco Group, Rotterdam. We thank the Robeco Group for giving us the permission to report the present results. Correspondence to Ph.H. Franses, Econometric Institute H11-15, Erasmus University Rotterdam, P.O.Box 1738 NL-3000 DR Rotterdam, The Netherlands

1 Introduction

With the advent of improved data collection techniques, the applied econometrician can nowadays have access to very large databases, in which the behavior of economic agents is recorded with substantial precision. In marketing, for example, scanner methods in supermarkets, customer loyalty cards and regular direct mailing all yield a wealth of information on current and potential customers, see Balasubramanian et al. (1998) for a recent survey. As computational limits recede, there are ample opportunities for a detailed econometric analysis of these large databases, although it appears that not all standard techniques that have been designed for small samples can be applied without further modification, see Granger (1998).

In some respects these large data bases have only a fairly low informational content. For example, in direct marketing it is quite common that the fraction of individuals who respond to a mailing is very small. A typical response rate in an untargeted direct mail campaign, where addressees are invited to donate or spend money, is .5%, so that out of 100000 individuals only 500 respond. Naturally, the subsequent econometric analysis often focuses on finding the characteristics that distinguish these 500 individuals from the mass of non-respondents.

Given a large fraction of non-respondents, one is tempted to omit the larger part of non-respondents from the analysis. This turns a random sample into a state-dependent sample, and hence one should adapt the econometric inference techniques to this new situation. In the present paper, we put forward such an adaptation for a censored regression model, known as the Tobit model of type I, see Amemiya (1985) for a survey. We choose this model as it is particularly relevant in marketing research, where the few respondents pay money for new products or donate to charity. It is then of interest to consider the sum spent in conjunction with the response rate.

The present capacity of desk computers does not call for a reduction of the data set. In fact, computing iterative Maximum Likelihood [ML] estimates for samples of thousands of observations is no longer particularly onerous. The true interest of the method lies in the fact that the data collection itself may be time-consuming and/or expensive. Marketing analysis that makes use of censored regression may typically combine data from three different sources: data bases of household characteristics (for example, based on zip codes), past individual purchase behavior, and responses from a current survey. In the example given below, we could have spared the effort of collecting nine-tenth of the information.

The outline of this paper is as follows. In Section 2, we summarize the main aspects of the censored regression model. In Section 3, we discuss ML parameter estimation in the case of endogenous sample selection. In Section 4, we provide some details on the large data set we subsequently use in a logit and Tobit analysis. The data concern about 9600 customers of a Dutch financial institution who hold savings accounts, and who may or may not shift part of their investment to equity funds during a period of time. In total, this renders about 300000 observations. The occurrence of the shift and the amount of money shifted are the two variables of interest. In our example, the second variable is heavily censored as shifts occur only in .5% of all cases. Sections 5 and 6 report on a binary logit model for the response as such, where we address coefficient estimates in Section 5 and the variance of the estimates in Section 6. This last analysis sheds some light on the rate at which the variances increase when the sample size is deliberately reduced. In Section 7, we apply the censored regression techniques outlined in Section 3, where we again delete various fractions of the non-switchers (or, non-respondents). In Section 8, we conclude.

2 The censored regression model

In this section we make use of the terminology and notation of Amemiya (1985) to present the simplest censored regression model. This is the standard Tobit model due to Tobin (1958), and it reads as follows:

$$y_i^* = x_i^T \beta + u_i, \tag{1}$$

$$y_i = y_i^* \qquad if \;\; y_i^* > 0, \tag{2}$$

$$y_i = 0 \qquad if \;\; y_i^* \leq 0, \tag{3}$$

where y_i is the observed response, the x_i are K observed variables, with the first equal to unity, and the u_i are i.i.d. drawings from a Normal distribution with zero mean and variance σ^2. We denote the standard Normal probability function by $\Phi(.)$ and the corresponding density by $\phi(.)$. The model implies a probit model for response, that is

$$P_i = P(y_i \neq 0) = \Phi(x_i^T\beta/\sigma), \tag{4}$$

where P_i denotes the probability that individual i responds. For an observation i with $y_i = 0$ the likelihood is therefore

$$L_{i,0} = 1 - \Phi(x_i^T\beta/\sigma), \tag{5}$$

while for an observation with $y_i \neq 0$

$$L_{i,1} = \sigma^{-1}\phi[(y_i - x_i^T\beta)/\sigma]. \tag{6}$$

Upon taking logarithms and summing we obtain the sample log-likelihood function in the unknown parameters β and σ.

In practice, one usually considers the parameters $h = 1/\sigma$ and $\alpha = \beta/\sigma$ to increase the speed of convergence of the ML estimation routine. The likelihoods then become

$$L_{i,0} = 1 - \Phi(x_i^T\alpha) \tag{7}$$

and

$$L_{i,1} = h\phi[(y_i - x_i^T\alpha]. \tag{8}$$

3 Estimation from a reduced sample

When we have a large full sample of n observations containing a very small number of m respondents, the analysis may be restricted to a reduced sample that is obtained by omitting a large part of the non-respondents. This selective reduction amounts to an endogenous selection of observations, and hence the method of estimation must be adapted accordingly. The general case is treated in Manski and Lerman (1977).

In this section, we first demonstrate the general principle for the simple case of a bivariate discrete model. Subsequently, we consider the censored regression model.

First, consider the binary response model. In the population the probability of response denoted by $Y_i = 1$, conditional upon regressor characteristics x_i, is

$$P_i^* = P(Y_i = 1|x_i), \tag{9}$$

and the probability of no response is its complement. Assume that the full sample is a random sample from the population with sampling fraction α. The probabilities that individual i is in this sample and has outcomes 1 or 0 are then αP_i^* and $\alpha(1-P_i^*)$. If however the number of zero observations is reduced by admitting only a randomly selected fraction γ to the reduced sample, these probabilities are αP_i^* and $\gamma\alpha(1-P_i^*)$. The probability of observing $Y_i = 1$ in the reduced sample is then given by

$$\tilde{P}_i = \frac{\alpha P_i^*}{\alpha P_i^* + \gamma\alpha(1-P_i^*)} = \frac{P_i^*}{P_i^* + \gamma(1-P_i^*)} \tag{10}$$

and the probability of observing no response is

$$1 - \tilde{P}_i = \frac{\gamma(1-P_i^*)}{P_i^* + \gamma(1-P_i^*)}. \tag{11}$$

The same argument can be applied to the censored regression model. Instead of the probability P_i of (4), the probability of observing response in the reduced sample now equals

$$\tilde{P}_i = \frac{P_i}{P_i + \gamma(1-P_i)}. \tag{12}$$

The likelihood values change accordingly, that is, they are multiplied by the same terms as in (10) and (11). For an observation i with $y_i = 0$, the likelihood is

$$\tilde{L}_{i,0} = L_{i,0}\ \frac{\gamma}{P_i + \gamma(1-P_i)}, \tag{13}$$

and for observations with $y_i \neq 0$ it is

$$\tilde{L}_{i,1} = L_{i,1}\ \frac{1}{P_i + \gamma(1-P_i)}. \tag{14}$$

The same parameters can be estimated as before by maximizing the log-likelihood function. This function now has a slightly more complicated form as it involves the known reduction factor γ. Combining (13) with (7) and (14) with (8), the overall likelihood in the reduced sample with reduction factor γ is

$$L_{i,0} = \gamma(1 - \Phi_i)/[\Phi_i + \gamma(1 - \Phi_i)] \tag{15}$$

and

$$L_{i,1} = h\phi(e_i)/[\Phi_i + \gamma(1 - \Phi_i)] \tag{16}$$

where

$$e_i = y_i - x_i^T \alpha, \tag{17}$$

$$\Phi_i = \Phi(x_i^T \alpha). \tag{18}$$

Consequently, the log-likelihoods are

$$\log L_{i,0} = C(\gamma) + \log(1 - \Phi_i) - \log[\Phi_i - \gamma(1 - \Phi_i)], \tag{19}$$

and

$$\log L_{i,1} = C(\sqrt{2\pi}) + \log h - 1/2e_i^2 - \log[\Phi_i - \gamma(1 - \Phi_i)] \tag{20}$$

with $C(.)$ denoting constants involving γ and $\sqrt{2\pi}$.

Taking derivatives of the log-likelihood results in

$$q_{i,0;\alpha} = \frac{-\phi_i}{(1 - \Phi_i)(\Phi_i + \gamma(1 - \Phi_i))} x_i, \tag{21}$$

$$q_{i,0;h} = 0, \tag{22}$$

and

$$q_{i,1;\alpha} = [e_i - \frac{(1 - \gamma)\phi_i}{\Phi_i + \gamma(1 - \Phi_i)}] x_i, \tag{23}$$

$$q_{i,1;h} = 1/h - e_i y_i. \tag{24}$$

With estimates $\hat{\alpha}$ and $\hat{h}$, we obtain the initial parameters upon using

$$\beta = \alpha/h \quad \sigma = 1/h, \tag{25}$$

The covariance matrix of β and σ is obtained from the covariance matrix of α and h by pre- and post-multiplication by a matrix G. As an example, with three regressor variables this matrix is

$$G = \begin{bmatrix} 1/h & 0 & 0 & -\alpha_1/h^2 \\ 0 & 1/h & 0 & -\alpha_2/h^2 \\ 0 & 0 & 1/h & -\alpha_3/h^2 \\ 0 & 0 & 0 & -1/h^2 \end{bmatrix}. \tag{26}$$

4 An illustrative large data set

In the sequel we apply the above results to a large data set. We consider a logit model for the occurrence of a switch of an individual from saving to investing, and we consider a type I censored regression model for the amount of money involved in that switch. Our interest lies in verifying if the parameter estimates in the full sample and in the reduced samples are approximately similar, which they should be when we rely on the corrected log-likelihood function. Additionally, we examine the estimated variances of the parameters in order to assess the potential loss of precision.

The data set we consider was made available by a large Dutch financial institution which offers a huge variety of savings accounts and investment funds in bonds and shares to the public. It has a large number of customers, who subscribe to one or more of its products. The customers can easily shift their holdings from one of the firms' funds to another, and this is indeed one of the major attractions of the firm. Our sample refers to about 9600 clients with a particular type of savings account and no other accounts with the firm. Every month during the period from January 1992 until August 1997 constitutes an observation, but only if there is a positive balance in the savings account at the beginning of the month. The censored regression model bears on the shift during that month of (part of) the savings balance to other investment funds offered by the same firm. If such a shift occurs the client drops out of the sample, for he or she no longer has a savings account and nothing else.

This selection of the data results in 293880 observations. Because of the requirement of a nonzero savings account and no other account at the beginning of the month, many individual records are incomplete. These records may start later than January 1992, terminate before August 1997, or have gaps in between. This precludes a thorough panel analysis of the data set. Therefore, we have to treat the observations as if they constitute one large sample, even though this is not strictly correct. Individual effects are ignored and so is the clear infringement of the assumption of independence. It is however not easy to find large data sets that are available for academic purposes, and the present data will have to do for our experimental analysis below.

Among the 293880 observations of the full sample only 1488 concern a shift from savings to another investment. Hence, there are 292392 zero (or non-respondent) observations. There is wide scope for reducing their number, and we consider this reduction for a range of values towards 1488.

We have five explanatory variables. This set has been defined by conventional wisdom and experience. Some variables refer to customer characteristics, while others concern events occurring in the month under review. The variables are (1) duration of customer relationship (Loyalty), (2) the (logarithm of) balance held in a savings account (Log Savings), (3) stock market return in the previous month (Share Index), (4) change in the savings interest rate in the previous month (Interest), and (5) a dummy variable for the introduction of new product in that month (New Product).

We estimate the parameters of the logit and Tobit models using ML for the full sample of 293880 observations and for several reduced samples. These reduced samples are constructed by adding random samples of non-respondents to the same 1488 respondent observations, every time taking multiples of 1488. The smallest reduced sample is evenly composed of the 1488 nonzero observations and 1488 zeros, with a reduction factor γ of $1488/(293880 - 1488) = .005089$. In the other samples the number of non-respondents is successively increased towards the full sample for which γ equals 1. These samples each contain K times 1488 non-respondents. The full range of K we consider is

$$1,\ 2,\ 5,\ 10,\ 25,\ 50,\ 100,\ 150,\ 196.5,$$

where the last value corresponds to the full sample. Each random sample of K times

1488 non-respondent observations is freshly drawn, but the resulting samples are not necessarily disjoint. Moreover, all samples used for estimation have the same 1488 respondent observations in common. Therefore, the various samples for parameter estimation are not independent.

ML estimation is carried out using the MaxLik library of GAUSS. We find that even the full sample of nearly 300000 observations can be handled on an ordinary desk computer with a 350 MHz Pentium II processor. The only special feature is that additional working memory is needed up to a total of 320 MB. The maximum working space in the GAUSS configuration file can then be set at 300 MB without necessitating swaps to and from the hard disk. This is sufficient to estimate a logit and a Tobit model with five regressor variables on the full sample.

5 A Logit Analysis: Coefficient Estimates

We first consider a logit model for the switch from saving to investing. We could consider a probit model as this corresponds with part of the Tobit model. For convenience we apply a logit model, because the specification of P_i^* in a logit model leads to a particularly simple solution for the estimation from a reduced sample using the method of Section 3. If P^* is a logit, $\tilde{P}$ and P^* differ only in the intercept and have the same slope coefficients. These slope coefficients can therefore be estimated directly by fitting a simple logit to the reduced sample, without any adjustment, and then retrieving the correct intercept of P^* by a simple adjustment of the intercept of $\tilde{P}$ by $\log(\gamma)$. This attractive method is discussed in for example Palepu (1986) and Hosmer and Lemeshow (1989).

The full sample estimates set the standard against which the other estimates must be judged. These are displayed in the final column of Table 1. Note once more that in the full sample the overall frequency of switching is only one half of a percent. This low frequency shows up in the intercept, but it does not affect the slope coefficients. All estimated parameters have the expected sign and high $t-$values. The negative effect of Loyalty (that is, the duration of the customer relationship without a change in the nature of this relation) should be interpreted as that inert customers, who did not change their behavior for a long time, are unlikely to do so now.

The same coefficients are estimated for a series of reduced samples of $(K+1)$ times 1488 observations each, that is, 1488 customers who change behavior and $K \times 1488$ zero observations. The estimation results are also given in Table 1. Table 2 shows the course of the six coefficients towards the full sample estimate. The estimates still show variation, in spite of the mutual dependence of the samples. The extent of this sampling variation must reflect the sample dispersion of the regressor variable concerned among the non-respondent observations, and hence the variation between successive selected samples. This differs from one regressor to another and hence from one coefficient to another. Differences in the overall dispersion of regressors are of course reflected by the standard errors of the full sample estimates. In Table 2 the deviations of each estimate from its full sample value are therefore reduced to a common denominator by expressing them in this full sample standard error.

We observe from Table 2 that the deviations from the full sample estimates steadily decline as K increases (as they should). For $K = 1$ there is only one coefficient with more than one standard error deviation, for $K = 2$ and $K = 5$ there are none, but for $K = 10$ we suddenly have two. For K beyond 25 the deviations are quite negligible. We hesitate to draw conclusions about an acceptable value of K from this fragmentary evidence, which after all reflects a single experiment. We postpone making this choice until we have considered the course of the variance of the estimates in the next section.

6 A Logit Analysis: Variance of the Estimates

As for the precision of the estimates, it is clear that any reduction of the sample involves a loss of information. Estimates from reduced samples are always less efficient than from a full sample. However, as the composition of the sample varies systematically with its size, this loss can be much smaller than the reduction in sample size suggests.

We investigate this issue by examining the course of the variances of the estimated slope coefficients of the logit model for various sample sizes. This variance is the inverse of the information matrix H, and it is given by

$$H = \sum \tilde{P}_i(1 - \tilde{P}_i)x_i^T x_i \tag{27}$$

with summation over the reduced sample. The derivation of the information matrix of a standard logit can be found in Cramer (1991, p.24), and (27) follows in the same manner from (12) upon substitution of the logit for P_i^*. We consider a single (diagonal) element of H and write the sum of (27) as the product of the number of terms n and a mean, which in turn is approximated by the product of two means, s in $H_j = n.mean(\tilde{P}_j(1-\tilde{P}_j)).mean(x_{ij}^2)$. We shall now examine the three elements in turn, that is, (1) the number of elements in the summation, which equals the sample size $(K+1)m$, with m the number of nonzero observations, (2) the observation weights $\tilde{P}_i(1-\tilde{P}_i)$, which for a given sample size enter by their sample mean, and (3) the specific regressor terms $x_i^T x_i$, also represented by their sample mean, that is, by the sample variance of x_j, $\sigma^2_{X(j)}$. Admittedly, this decomposition ignores any systematic relations between the weights and the regressor values, and also any correlation among regressor variables, that is, the off-diagonal elements of $x_i^T x_i$. We assume that these effects do not vary with the size (and composition) of the sample.

We can approximate each of the three elements as a function of K, and together these terms determine the course of H_j and of the variance V_j as a function of K. In varying K, we examine the effect of adding successive batches of zero observations to a given number of respondent observations rather than a progressive reduction of the full sample. The starting point is the minimum reduced sample with $K = 1$, with a total number of $2K$ observations, which we will call the basic sample. Below, we use the statistics from this sample.

As already noted, the number of terms in (27) is $(K+1)m$. Hence, the sample size effect is

$$H_1 = (K+1)m. \tag{28}$$

The observation weight $\tilde{P}_i(1-\tilde{P}_i)$ obtains it maximum at $\tilde{P} = .5$, and it tails off as $\tilde{P}$ moves away from this value. In each sample, the $\tilde{P}_i$ are centered around their mean, which is equal to $K/(K+1)$. If the sample dispersion is ignored, the mean observation weight would be $K/(K+1)^2$. Clearly, the minimum reduced sample has the most advantageous weights $\tilde{P}(1-\tilde{P})$, with this factor being equal to .25. Adding another m zero observations reduces it to .11, and from then on it declines more slowly to zero. Closer inspection, however, shows that this first approximation overstates the true sample mean, in particular if $\bar{P}$ is close to .5 for then the weight declines on both sides of this central point. A better approximation is found by incorporating the sample dispersion of the $\tilde{P}_i$ by taking the mean of a second order Taylor expansion. This gives

$$H_2 = \bar{P}(1-\bar{P}) - \sigma^2_{P(K)}, \tag{29}$$

with $\sigma^2_{P(K)}$ the sample variance of the $\hat{P}_i$ for a given value of K. This can be assessed in turn by the approximation

$$\sigma^2_{P(K)} \approx \left(\frac{dP(K)}{dP(1)}\right)^2 \sigma^2_{P(1)}, \tag{30}$$

which gives

$$\sigma^2_{P(K)} \approx \frac{16K^2}{(K+1)^4}\sigma^2_{P(1)}. \tag{31}$$

Upon substituting this result, as well as $\bar{P} = 1/(K+1)$, into (29) we obtain

$$H_2 = \frac{K}{(K+1)^2}C(K) \tag{32}$$

with the correction term $C(K)$ given by

$$C(K) = 1 - \frac{16K}{(K+1)^2}\sigma^2_{P(1)}. \tag{33}$$

This term is always less than 1 and approaches 1 as K increases.

H_2 is a powerful correction of the size effect H_1. This is easily seen from multiplying H_1 with H_2, which apart from the scale factor m results in $(K/(K+1))C(K)$. This factor rises quickly from a value slightly below .5 to an upper bound of 1 as K increases. Hence, increasing the sample size selectively by adding non-respondent (zero) observations while keeping the number of respondents constant, does not reduce the variances towards zero, but towards a finite limit.

The third term is specific for each particular slope coefficient as it concerns the variance of X_j. This variance can be decomposed for the two sub-samples of respondents and non-respondents into within-group variation and between-group variation. The between-group variance can be written in terms of the difference between the two group means, that is, $\Delta X_j(K) = \bar{X}_{1,j}(K) - \bar{X}_{0,j}(K)$. K intervenes in the sub-sample weights of the two groups $1/(K+1)$ and $K/(K+1)$. Altogether we have

$$\begin{aligned} H_3 &= \sigma^2_{X(j),K} \\ &= 1/(K+1)\sigma^2_{X(j),0,K} + K/(K+1)\sigma^2_{X(j),1,K} \\ &\quad + K/(K+1)^2(\Delta X_j(K))^2. \end{aligned} \tag{34}$$

In practice, we equate the three constants involved to their values for $K = 1$. Like H_2, this term declines with increasing K, and thus mitigates the pure size effect. Its sensitivity to K depends on the ratio of ΔX_j^2 to $\sigma_{X(j),1}$, which reflects the discriminating power of $X(j)$ between $Y = 1$ and $Y = 0$, or its importance as a determinant of the binary choice. The more important a regressor, the greater the decline of H_3, the smaller the overall increase of the total H, and the smaller the loss of efficiency through reducing the sample size selectively.

Upon multiplying H_1, H_2 and H_3 from (28), (32) and (35) we obtain

$$H_j \approx \frac{K}{K+1} C(K) \sigma^2_{X(j),K}.m \tag{35}$$

As we neglect the off-diagonal elements of the matrix H, the extrapolated value of the variance of the estimated coefficient is obtained by taking the inverse of this element, that is,

$$V_j = H_j^{-1}. \tag{36}$$

The four constants involved in $C(K$ and $\sigma^2_{X(j),K}$ are $\sigma^2_{P(1)}$, $\sigma^2_{X(j),1,1}$, $\sigma^2_{X(j),0,1}$ and $\Delta X_j(1)$, and these can all be derived readily from the analysis of the single reduced sample for $K = 1$. The first three terms of (35) are the same for all coefficients. The last one is specific for each regressor. They are given in Table 3.

As already noted, H does not increase beyond all limits with K. Selective increases in sample size do not reduce the variances towards zero, but towards a finite limit. From (35) we find for $K \to \infty$ that

$$\bar{H}_j = \sigma^2_{X(j),0,1} \tag{37}$$

This corresponds with a minimum value of the variances given by

$$\bar{V}_j = \bar{H}_j^{-1}, \tag{38}$$

also given in Table 3.

For the present data, we obtain the results as presented in Table 4. This table gives the calculated or extrapolated variance of each slope coefficient of the logit model and (in italics) the actual value for the full range of K. All have been expressed in terms of their lower limit from (38). These minimum variances are given in Table 3, along with the regressor-specific constants used in the calculation of the extrapolated variances. In our case, $\sigma^2_{P(1)}$ has the value .0314.

A comparison with the implied t-values of the estimates in Table 1 shows that the ratio of between-group variance to within-group variance indeed roughly reflects the explanatory power of the regressors.

We see in Table 4 that for some regressors the actual variances systematically exceed the extrapolated values, although only by about 10 to 25 percent. This must be attributed to a slight inter-correlation between the regressors, that is, between loyalty and log savings throughout all samples, and between share index and interest for non-respondents only. This explains why the latter effect occurs more strongly as K increases and the proportion of non-respondent observations rises.

On the whole, the extrapolated variances faithfully reproduce the course of the variances. The first main conclusion from Table 4 is then that even in the basic sample, which contains only about 3000 observations from the available total of about 300000, the variances are at most 2.6 times as large as the minimum value (which is almost attained by the full sample). This implies that by the strongly selective nature of the manner of sample reduction, we have a reduction in efficiency only by a factor .4 instead of by .01. The second main conclusion is that if one accepts a loss of efficiency of, say, twenty percent from the results from the full sample, a value of K of 10 will by and large be quite sufficient.

7 A censored regression analysis

In this section we report on fitting the Tobit model to the same set of samples, first estimating the parameters $h = 1/\sigma$ and $\alpha = \beta/\sigma$ of (7) and (8) and then converting these to the original parameters β and σ. Again the full sample estimates, shown in the final column of Table 5, set the standard for the other estimates. We recall that the sample is vast but the overall sample frequency of response is minute, that is, 1488 nonzero observations against 292392 zeros. The observations of this tiny minority are however more informative than the others, for they report a value of the latent variable while a zero observation only reports its sign. As before, the coefficients are estimated with considerable precision, see Table 5.

The same coefficients have been estimated from the same series of reduced samples of $(K + 1)$ times 1488 observations as before, and the other columns in Table 5 display the estimation results. We observe that the point estimates are very close for all samples, and that the variances show a slow decrease with increasing values of K.

This is confirmed by Table 6, which shows the course of the six coefficients towards their final estimate. As before, the deviations of each estimate from its final value have been expressed in their final standard error. For $K = 1$, three out of the seven estimates deviate more than one standard error from the norm. The deviations decline slightly more quickly than in the case of the logit, but the difference is small. From $K = 10$ onwards the parameter estimates differ hardly from the full sample estimates. Hence, with a sample size of 16000 instead of 300000, we have almost the same result even though we use only one twentieth of the data set.

Finally, Table 7 shows the behaviour of the variances of the β parameters as a function of sample size. Note that the variance of the estimated σ is a function of σ itself, and hence of not much interest. The variances have been normalized by expressing each in the variance for the full sample. The overall behaviour is quite similar to the outcomes for the logit model, although we should stress that the algebra that we invoked in that case to explain the effect of K does not apply here.

8 Concluding remarks

In this paper we derived the necessary statistical theory for analyzing a censored regression model, where we deliberately deleted (large numbers of) non-respondents. We showed that the theory works in a practical situation, and also that the loss of precision does not have to be large.

Our results should be useful to applied econometricians who have to analyze data sets with only few informative (or interesting) observations and a potentially very large number of non-respondent (or non-informative) observations. We showed that one does not need to collect all data (at possibly high costs and with substantial efforts), and that only a fraction of the non-respondents will do.

K	1	2	5	10	25	50	100	150	196.5
Observations	2976	4464	8928	16368	38688	75888	150288	224699	293880
Constant	-4.848 (.120)	-4.940 (.103)	-4.861 (.094)	-4.810 (.087)	-4.840 (.083)	-4.875 (.082)	-4.877 (.081)	-4.874 (.080)	-4.874 (.080)
Loyalty	-.287 (.100)	-.242 (.084)	-.208 (.075)	-.295 (.073)	-.271 (.069)	-.260 (.068)	-.269 (.068)	-.261 (.067)	-.264 (.067)
Log Savings	.470 (.035)	.443 (.030)	.463 (.027)	.476 (.025)	.473 (.023)	.459 (.022)	.461 (.022)	.460 (.022)	.460 (.022)
Share Index	1.388 (.528)	1.699 (.455)	1.185 (.393)	1.411 (.378)	1.487 (.366)	1.498 (.358)	1.557 (.354)	1.533 (.354)	1.514 (.353)
Interest	-1.198 (.156)	-1.224 (.139)	-1.140 (.126)	-1.087 (.118)	-1.198 (.117)	-1.198 (.116)	-1.220 (.116)	-1.212 (.116)	-1.208 (.116)
New product	.846 (.084)	.767 (.070)	.756 (.061)	.815 (.057)	.751 (.055)	.752 (.054)	.766 (.053)	.755 (.053)	.755 (.053)

Table 1. Parameter estimates for a binary logit model for a switch from saving to investment, obtained for the full sample (final column) and for various reduced samples (the intercept coefficient has been appropriately adjusted). Standard errors in parentheses.

K	1	2	5	10	25	50	100	150
Observations	2976	4464	8928	16368	38688	75888	150288	224699
Constant	32	-83	16	80	43	-1.3	-4.0	-4.0
Loyalty	-35	32	84	-47	-10	5.1	-7.6	3.7
Log Savings	46	-78	13	75	59	-3.2	6.5	1.4
Share Index	-36	52	-93	-30	-7.7	-4.5	12	5.4
Interest	9.0	-13	59	104	8.5	8.9	-10	-3.5
New product	172	23	1.3	114	-7.2	-5.1	-55	0.4

Table 2. Logit parameter estimates for various reduced samples, expressed as deviations from the full sample values in per cent of the standard errors of the latter estimates.

	$\sigma^2_{X(j),1}$	$\sigma^2_{X(j),0}$	$(\Delta X_j(i))^2$	$\bar{V}_j$
Loyalty	.1798	.1655	.0099	.0041
Log Savings	1.4444	1.3892	.6692	.0005
Share Index	.0058	.0054	.0023	.1242
Interest	.0723	.0553	-.0765	.0122
New Product	.2462	.1898	.1835	.0035

Table 3. Some constants for the basic reduced sample ($K = 1$, $n = 2976$)

K	1	2	5	10	25	50	100	150	196.5
Loyalty	2.19	1.64	1.27	1.14	1.06	1.03	1.01	1.01	1.00
	2.44	*1.74*	*1.39*	*1.30*	*1.17*	*1.14*	*1.12*	*1.11*	*1.10*
Log Savings	2.12	1.58	1.23	1.11	1.05	1.02	1.01	1.01	1.01
	2.59	*1.86*	*1.50*	*1.28*	*1.13*	*1.04*	*.99*	*.98*	*.96*
Share Index	2.21	1.65	1.28	1.14	1.06	1.03	1.01	1.01	1.01
	2.24	*1.66*	*1.24*	*1.15*	*1.08*	*1.03*	*1.01*	*1.01*	*1.01*
Interest	1.94	1.50	1.21	1.11	1.04	1.02	1.01	1.01	1.01
	2.01	*1.58*	*1.11*	*1.15*	*1.13*	*1.10*	*1.11*	*1.11*	*1.10*
New Product	1.92	1.48	1.20	1.10	1.04	1.02	1.01	1.01	1.01
	1.98	*1.39*	*1.05*	*.93*	*.84*	*.82*	*.80*	*.80*	*.79*

Table 4. Extrapolated and actual variances (in italics) of the slope coefficients in the binary logit model for switching from saving to investing. The variances have been expressed in multiples of their limiting value of Table 3.

K	1	2	5	10	25	50	100	150	196.5
Observations	2976	4464	8928	16368	38688	75888	150288	224699	293880
Constant	-1.344 (.028)	-1.379 (.025)	-1.368 (.023)	-1.364 (.022)	-1.370 (.022)	-1.381 (.022)	-1.387 (.022)	-1.387 (.022)	-1.387 (.022)
Loyalty	-.050 (.019)	-.044 (.016)	-.040 (.014)	-.057 (.014)	-.053 (.013)	-.051 (.013)	-.056 (.013)	-.054 (.013)	-.054 (.013)
Log Savings	.115 (.006)	.106 (.005)	.107 (.005)	.108 (.004)	.108 (.004)	.105 (.004)	.105 (.004)	.105 (.004)	.104 (.004)
Share Index	.293 (.106)	.342 (.092)	.239 (.081)	.283 (.077)	.298 (.074)	.300 (.073)	.305 (.073)	.299 (.073)	.295 (.073)
Interest	-.242 (.031)	-.248 (.027)	-.233 (.023)	-.243 (.022)	-.245 (.021)	-.246 (.021)	-.254 (.021)	-.253 (.020)	-.252 (.020)
new product	.154 (.017)	.245 (.015)	.147 (.013)	.159 (.012)	.148 (.012)	.148 (.012)	.156 (.012)	.154 (.012)	.154 (.012)
σ	.573 (.005)	.576 (.005)	.576 (.015)	.577 (.005)	.577 (.005)	.578 (.005)	.579 (.005)	.580 (.005)	.580 (.005)

Table 5. Estimates for the type I censored regression for the amount of money switched from saving to investing. Standard errors in parentheses.

K	1	2	5	10	25	50	100	150
Observations	2976	4464	8928	16368	38688	75888	150288	224699
Constant	198	38	87	104	78	29	0.5	.05
Loyalty	32	80	114	-16	11	26	-7.9	3.2
Log Savings	245	29	70	92	74	14	7.0	2.7
Share Index	-1.8	65	-76	-16	5.2	7.7	14	6.6
Interest	50.5	21	97	139	36	31	-8.4	-2.2
New Product	-.05	-77	-61	46	-50	-48	19	-.6
σ	-122	-66	-65	-50	-50	-23	-1.9	-3.2

Table 6. Parameter estimates for a censored regression model for various reduced samples, expressed as deviation from the full sample values and scaled by the standard errors of the estimates for the full sample ($\times 100$).

K	1	2	5	10	25	50	100	150
Observations	2976	4464	8928	16368	38688	75888	150288	224699
Loyalty	2.142	1.604	1.249	1.130	1.041	1.018	1.006	1.000
Log Savings	2.056	1.556	1.222	1.111	1.056	1.056	1.000	1.000
Share Index	2.114	1.610	1.250	1.135	1.046	1.019	1.008	1.002
Interest	2.293	1.740	1.320	1.154	1.058	1.024	1.010	1.002
New Product	2.246	1.627	1.261	1.142	1.045	1.015	1.007	1.007

Table 7. Actual variances of the slope coefficients in the censored regression for various reduced samples, expressed as multiples of the full sample values.

References

Amemiya, T. (1985), *Advanced Econometrics.* Harvard University Press, Cambridge, Mass.

Balasubramanian, S., S. Gupta, W. Kamakura and M. Wedel (1998), Modelling Large Data Sets in Marketing. *Statistica Neerlandica* **52**, 303-323.

Cramer, J.S. (1991), *The Logit Model - An Introduction for Economists.* London: Arnold.

Granger, C.W.J. (1998), Extracting Information from Mega-panels and High-frequency Data. *Statistica Neerlandica* **52**, 258-272.

Hosmer, D.W., and S. Lemeshow (1989), *Applied Logistic Regression.* New York: Wiley.

Manski, C.F., and S.R. Lerman (1977), The Estimation of Choice Probabilities from Choice-based Samples. *Econometrica* **45**, 1977-1988.

Palepu, K.G. (1986), Predicting Takeover Targets, *Journal of Accounting and Economics* **8**, 3-25.

Tobin, J. (1958), Estimation of Relationships for Limited Dependent Variables. *Econometrica* **26**, 24-36.