Abstract

This thesis contributes in several ways to the existing knowledge on estimation of truncated, censored, and left truncated right censored (LTRC) regression models. Three new semiparametric estimators are proposed, allowing for asymmetric error distributions. A bootstrap method for estimation of the covariance matrix of the quadratic mode estimator (QME) is proposed and studied. In addition, finite sample properties of estimators for truncated, censored, and LTRC data are studied within simulation studies and applications with real data.

The first paper consists of a simulation study of the QME and other estimators of truncated regression models. The paper contributes with results suggesting the bootstrap technique being potentially useful for estimation of the QME covariance matrix.

In the second paper estimators of truncated and censored semiparametric regression models are proposed. These estimators are generalizations of the QME and the winsorized mean estimator (WME) by allowing asymmetric "trimming" of observations. Consistency and asymptotic normality of the estimators are shown.

By combining the two moment restrictions used to derive the estimators in the second paper, a consistent estimator of LTRC regression models is proposed in the third paper.

The fourth paper contains an application where LTRC interpurchase intervals of cars are analysed. Results regarding the interpurchase behaviour of consumers are provided, as are results on estimator properties.

Keywords and phrases: LTRC, consistency, asymptotic normality, bootstrap, interpurchase intervals.

AMS 2000 subject classification: 62J99, 62F12, 62F40.

Acknowledgements

Who would have guessed 11 years ago when I began my university studies that I would write a doctoral thesis? I certainly did not but now I am here, writing the last words in the thesis... It has been a lot of fun but also a lot of hard work, many laughs and a few tears during the years writing this thesis. I have received great help and support from many persons. Without your help this thesis would not have been finished. I would like to thank everyone who has supported me. Some of you I wish to give special thanks to.

My supervisor, Associate Professor Thomas Laitila, who receives huge thanks. You have read numerous earlier drafts of the papers in this thesis and your comments and criticism have improved them substantially. Thank you for sharing your knowledge and experience of research with me. Most valuable to me has been that you never failed to find encouraging words when I doubted my ability.

My family, Jörgen and Albin, you are my biggest dream come true. Thank you for all your love and great support and for reminding me about what is really important in life. Jörgen, thank you for having patience with my bad temper the last months when I have been under stress about work and leaving Albin at "förskolan" and so on.

My parents, Astrid and Rolf, and my sisters, Emma and Jenny, for love and support. Mum and Dad, thanks for enabling me to work longer hours the last few weeks by baby-sitting Albin and by helping me with other practical things. I really appreciate your help.

Professor Göran Broström and Associate Professor Xavier de Luna for reading and discussing earlier versions of some of the papers at a seminar.

My friend and colleague, Dr. Erling Häggström Lundevaller, for helping me in many ways during the years by discussing issues in

statistics and programming, being my “shrink”, and making me laugh.

My colleagues (present and former) at the Department of Statistics for sharing your knowledge with me and for your friendship.

And last but not least, special thanks to my friends and relatives for your friendship and support and for giving me a bit of a breathing-space from work.

Umeå, May 2005

Maria

Contents

1 Introduction **1**

2 Truncated and censored regression models **2**
2.1 Truncation and censoring 5
2.2 ML estimation under truncation and censoring 8
2.3 Semiparametric models 10
2.4 Semiparametric estimators 11

3 Summary of papers **14**
3.1 Paper I: Finite sample properties of the QME 15
3.2 Paper II: Estimators of regression parameters for truncated and censored data 17
3.3 Paper III: A semiparametric regression estimator under left truncation and right censoring 18
3.4 Paper IV: Analysis of interpurchase intervals of cars using truncated and censored data 19

4 Further research **21**

List of papers

I. Karlsson, M. (2004). Finite sample properties of the QME. *Communications in Statistics - Simulation and Computation*, **33**, 567-583.

II. Karlsson, M. (2005). Estimators of regression parameters for truncated and censored data. *Metrika* (In press).

III. Karlsson, M., Laitila, T. (2005). A semiparametric regression estimator under left truncation and right censoring. Research Report 2005:5, Department of Statistics, Umeå University, Umeå.

IV. Karlsson, M. (2005). Analysis of interpurchase intervals of cars using truncated and censored data. Research Report 2005:6, Department of Statistics, Umeå University, Umeå.

I am indebted to Marcell-Dekker, Inc. for kind permission to include Paper I and to Springer Science and Business Media for kind permission to include Paper II.

1 Introduction

Statistical methods are applied in most areas of research and for a variety of purposes. For dimensioning of logistic systems, statistics are employed to predict demand. When developing new medical drugs, statistics are used for evaluation of their properties. In theoretical research, theories are tested by comparison with empirical observations using statistics. Thus, statistics and statistical research methods are important for everyday life and the progress of scientific knowledge.

A vast flora of different standard statistical methods is made available to practitioners in statistical computer software packages. These methods rest upon assumptions about the data used in the analysis and the population from which the data is generated. However, there are many situations when these assumptions are not fulfilled whereby standard methods of analysis may be inappropriate. For these situations modifications of available methods are required as is the development of new methods.

In this thesis estimation of linear regression models using truncated and censored data is considered. Linear regression analysis is one of the most common methods for analysis of relationship between two or more variables and there are well-known methods for estimation of regression models available. However, it is known that these standard methods for estimation of regression models are not suitable when data is truncated or censored, because then the estimators are biased and inconsistent. This means that the estimators either systematically overestimate or underestimate the unknown parameters and furthermore, they do not even converge to the true parameter values when the sample size increases.

Several suggestions for alternative estimators have been made. Many of them are estimators of so called semiparametric regression models, i.e., models with only weak regularity conditions placed on

the random part of the model. This thesis contributes with proposals of additional estimators of semiparametric truncated and censored regression models. One interesting feature of the estimators proposed here is that they can be employed under asymmetric distributions, a feature not shared with many existing estimators. The estimators therefore provide alternatives when symmetry of the distribution is not plausible.

This thesis also provides new studies of the properties of estimators for truncated and censored regression models. Results are derived from a number of simulation studies as well as from empirical applications. Some of the simulation studies are also based on real empirical data. In addition, a proposal for using the bootstrap technique for estimation of the covariance matrix of one of the existing estimators is given. The performance of the bootstrap estimator is studied in a simulation study. Finally, results from an empirical study of interpurchase intervals of cars are presented. Earlier suggestions for parametric models of interpurchase intervals are evaluated, and the effects of individual demographic characteristics as explanatory variables are estimated.

In addition to this summary, the thesis consists of Papers I–IV. The outline of the summary is as follows: In Section 2 linear regression models are introduced along with truncated and censored regression models. Estimation of the models is also considered. Section 3 gives a summary of Papers I-IV, and the last section presents some ideas for further research.

2 Truncated and censored regression models

A linear regression model is used to study the relationship between one response (dependent) variable and one or several explanatory (independent) variables. For example, the response variable can be

nutritional content in boiled potatoes and the explanatory variables can be boiling time and whether the potatoes are peeled off or not; the response can be household expenditure on food and the explanatory variables can be household size and income; the response can be circumference of fish and the explanatory variables might be length, weight, and species of the fish.

The form of a linear regression model is

$$Y_i = X_i^T \beta_0 + \varepsilon_i, \qquad i = 1, 2, ..., n, \tag{1}$$

where Y_i is the response variable, X_i is a p-dimensional vector of explanatory variables, β_0 is a p-dimensional vector of unknown parameters, and ε_i is the random error term or disturbance. The error terms ε_i $(i = 1, \cdots, n)$ are assumed to be independent and identically distributed (i.i.d.) with mean 0 and variance σ^2.

The objective with regression analysis is to estimate the unknown parameter vector β_0 using a sample of observations of the response variable Y_i and the explanatory variables X_i. After estimating the parameters, the model can be used for a number of different purposes. One purpose might be to validate the theoretical assumptions on the relationship between the response and explanatory variable. For example, does the nutritional content in boiled potatoes tend to decrease with the boiling time? Another purpose might be to use the model for prediction, i.e., to use known values on the explanatory variables to predict the value of the response variable. For example, based on family size and family income, the expenditure on food might be predicted.

There are well-known methods for estimation of regression models available in most statistical computer software packages. The so-called ordinary least square (OLS) and maximum likelihood (ML) estimators are the most commonly adopted estimators.

The OLS estimator is defined as the parameter vector β which

minimises the sum of squared errors, i.e.,

$$\hat{\beta}_{OLS} = \arg\min_{\beta \in B} \sum_{i=1}^{n} \left(Y_i - X_i^T \beta\right)^2 \tag{2}$$

The solution to (2) is $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$, where X is the $n \times p$ matrix with the p-dimensional vectors X_i as rows and y is the $n \times 1$ vector of the n observations of the response variable.

The OLS estimator has desirable properties as an estimator of parameters of (1) under some regularity conditions on the model. Some of the more important conditions, besides that the linear functional form of the relationship in (1) holds, are that X has full column rank and that $E(\varepsilon|X) = 0$. Then the OLS estimator is unbiased, i.e., the expected value of the OLS estimator equals the true parameter vector β_0.

Under some additional assumptions, the OLS estimator is a consistent estimator of β_0, i.e., $\hat{\beta}_{OLS}$ converges in probability to the true parameter vector β_0. The OLS estimator is asymptotically normally distributed, i.e., $\sqrt{n}(\hat{\beta}_{OLS} - \beta_0)$ converges in distribution to a normal distribution with mean vector 0 and covariance matrix $\sigma^2 Q^{-1}$, where $Q = \text{plim}_{n\to\infty} 1/n X^T X$. If, in addition, the error terms are normally distributed, then the sampling distribution of the OLS estimator conditional on X is the normal distribution with mean vector β_0 and covariance matrix $\sigma^2 (X^T X)^{-1}$. (Details on assumptions for these results see, for example, theorems 2.28 and 4.25 in White (2001).)

Normality allows inference about the parameters by construction of confidence intervals and testing hypotheses using the t- and F-distributions. Without normally distributed error terms, the finite sample distribution of the estimator is unknown. However, if asymptotic normality is established, the sampling distribution in finite samples is approximated with the normal distribution implied

by the asymptotic distribution.

The objective with ML estimation is to find parameter estimates, $\hat{\theta}_{ML}$, such that the likelihood function $L(\theta|y, X) = \prod_{i=1}^{n} f(y_i, X_i|\theta)$, i.e., the joint density of the sample as a function of the parameters θ, is maximised. The ML estimator $\hat{\theta}_{ML}$ is defined by

$$L(\hat{\theta}_{ML}|y, X) > L(\theta|y, X) \qquad \text{for all } \theta \in \Theta \text{ such that } \theta \neq \hat{\theta}_{ML}, \quad (3)$$

where $\theta = (\beta, \eta)$, and η are parameters in the density function $f(\cdot)$ of ε in (1). For example, if the error term in (1) is assumed to be normally distributed with mean zero and variance σ^2 then $\theta = (\beta, \sigma^2)$. In practice, it is usually the log-likelihood function, $\ell(\cdot) = \ln L(\cdot)$, that is maximised with respect to the parameters θ instead of the likelihood function itself, because the log function is monotonically increasing and easier to work with.

ML estimators can be shown consistent and asymptotic normal distributed under quite general conditions (see e.g., theorems 17 and 18 in Ferguson, 1996). Also, under normally distributed errors the OLS and ML estimators of β_0 are identical.

Unfortunately data is, for different reasons, sometimes not complete and this complicates estimation and inference. Then the standard estimation techniques may no longer have desirable properties such as unbiasedness and consistency. In this thesis estimation of regression models with a truncated and censored response variable is considered. Truncation and censoring can arise due to sampling from incomplete populations or limitations of the study design or measurement equipment.

2.1 Truncation and censoring

The regression model (1) with a truncated response variable, Y_i, is known as a truncated regression model. A left (right) truncated response variable means that observations of (Y_i, X_i) are obtained

only for the part of the population for which $Y_i > t_i$, ($Y_i < t_i$) where t_i is the truncation point. This is equivalent to $\varepsilon_i > t_i - X_i^T\beta$ ($\varepsilon_i < t_i - X_i^T\beta$) expressed in error terms. Left and right truncation are sometimes called truncation from below or above, respectively.

The regression model (1) with a censored response variable, Y_i, is called a censored regression model. A right (left) censored regression model means that if the value of the response is larger (smaller) than a censoring point s_i then the value s_i is observed instead of the true value of Y_i. It is known that the value of Y_i is larger (smaller) than s_i but not exactly what the value is. As for truncated data, the words left and right censoring can be replaced by censoring from below or above. The response variable is called doubly censored if the sample is both left and right censored.

The literature distinguishes between two ways of censoring. Type I censoring occurs when the censoring point is fixed in advance. A sample of n units is drawn and the observations falling above (or below) the predetermined limit are censored. Thus the number of censored observations is a random variable not known in advance. For Type II censoring, on the other hand, the number of censored observations is known (fixed) in advance but the censoring point is not. In this thesis Type I censoring is considered.

As an illustration of these problems, consider the following example. In a study of circumferences of fish a sample of fish is caught from a lake using a fishing net. The smallest fish swim through the net and are not included in the sample. Thus, all fish with a circumference smaller than the circumference of the holes in the fishing net are excluded from the sample. Thus, neither the circumference nor any other characteristics such as length and weight of these fish are measured. Neither is it known how many of these fish exist or even if they exist at all. The response, circumference of fish, is said to be left truncated at t, the circumference of the holes in the fishing net. Truncation can also be explained as sampling from an incom-

plete population. Fish are sampled from the population consisting of "fish with circumference larger than t".

In the example with circumference of fish, the response would be right censored at s if, for example, the tape measure used to measure the circumference of the captured fish is too short for the biggest fish. Then the censoring point s is the length of the tape measure. The censored circumferences exact length are unknown but are known to be longer than the length of the tape measure.

In analysis of durations it is very common that data is right censored. Duration analysis can for example be studies of failure times for electronic components in an industrial experiment, studies of survival time for mice in a medical experiment, or an unemployment spell in an economic study of vocational training for the unemployed. If the study is terminated before all durations have been completed then durations are right censored. For example, it is only known that an unemployment spell lasted longer than s days for those individuals still without employment when the study was terminated after s days.

If data is both left truncated and right censored it is called LTRC data. The example with circumference of fish data is LTRC. The smallest fish swam through the fishing net and the circumferences are left truncated at t. The largest fish circumferences are only know to be longer than the length, s, of the tape measure and are thus right censored.

In this thesis left truncation and right censoring and the combination of the two, LTRC, are considered. However, right truncated data can be transformed to left truncated data by multiplying by -1 and vice versa. The analogue is true for censored data. LTRC data can be transformed to RTLC data in a similar manner.

As previously implied, standard estimation techniques are not suitable for truncated and censored regression models. The OLS estimator (2) is usually biased and inconsistent and the ML estima-

tor (3) is sensitive to distribution misspecification (see e.g., Chapter 15 in Davidson and MacKinnon, 1993). There is a need for other estimators. Such estimators are considered in the next sections.

2.2 ML estimation under truncation and censoring

The OLS estimator (2) is a biased and inconsistent estimator of truncated and censored regression models because then $E(\varepsilon|X)$ is a function of X and not equal to zero. The ML estimator (3) can however be used. Consider model (1) under truncation, $Y_i > t_i$. The likelihood function to maximise with respect to (β, η) is then

$$L(\beta, \eta|y, X) = \prod_{i=1}^{n} \frac{f(y_i - X_i^T\beta|\eta)}{\big(1 - F(t_i - X_i^T\beta|\eta)\big)}, \tag{4}$$

where $f(\cdot)$ and $F(\cdot)$ denote the probability density function (pdf) and cumulative density function (cdf) of the error term, respectively. Thus, the likelihood (4) is the density of ε_i conditional on ε_i being included in the sample, i.e., that $\varepsilon_i > t_i - X_i^T\beta$.

Similarly, if the problem is censoring, $Y_i = \min(Y_i, s_i)$, instead of truncation, let z_i denote an indicator function which equals 1 if observation i is uncensored and 0 if it is censored. Then the likelihood function is

$$L(\beta, \eta|y, X) = \prod_{i=1}^{n} \big(f(y_i - X_i^T\beta|\eta)\big)^{z_i} \big(1 - F(s_i - X_i^T\beta|\eta)\big)^{(1-z_i)} \tag{5}$$

Uncensored observations contribute to the likelihood by the factor $f(y_i - X_i^T\beta|\eta)$ while censored observations contribute with the factor $1 - F(s_i - X_i^T\beta|\eta)$. Note that $1 - F(s_i - X_i^T\beta|\eta)$ is the probability that $\varepsilon_i > s_i - X_i^T\beta$, i.e., the probability that an observation is censored.

For LTRC data the likelihood function is

$$L(\beta,\eta|y,X)=\prod_{i=1}^{n}\frac{\left(f(y_i-X_i^T\beta|\eta)\right)^{z_i}\left(1-F(s_i-X_i^T\beta|\eta)\right)^{(1-z_i)}}{\left(1-F(t_i-X_i^T\beta|\eta)\right)} \tag{6}$$

The advantages of ML estimators are many. As stated above, under some regularity conditions ML estimators are consistent and asymptotic normally distributed. Furthermore, estimators of the covariance matrix of the ML estimator have been derived. Usually the Hessian, the outer-product-of-the-gradient (or BHHH estimator), or the information matrix evaluated at the estimated parameters (β,η) is used (e.g., Davidson and MacKinnon, 1993). Another major advantage of using ML is that there is a developed theory for statistical inference. Examples are testing hypotheses about the parameters using so called likelihood ratio tests and comparison of models through measures based on the log-likelihood value, e.g., the Akaike information criterion (AIC).

However, there are also some disadvantages with ML estimation. First the pdf $f(\cdot)$ used to formulate the likelihood function has to be decided on. For truncated and censored regression models the normal distribution is often used. The ML estimator of the censored regression model, i.e., the maximum of (5) with respect to (β,η) assuming normally distributed errors is the well-known Tobit ML estimator. The maximum of (4) with respect to (β,η) assuming normally distributed errors is sometimes also referred to as the Tobit ML estimator due to its relation to the censored regression model (Hausman and Wise, 1977).

Another disadvantage of the ML estimator is that it is sensitive to distributional misspecification (e.g., Davidson and MacKinnon, 1993). Thus, the decision on which pdf $f(\cdot)$ to use in the expressions (4), (5), and (6) matters for the properties of the estimator. Results in Vijverberg (1987) show the implications in terms of bias using the

Tobit ML estimator for censored and truncated regression models when the error distribution is non-normal.

2.3 Semiparametric models

Several suggestions for alternative estimators of truncated and censored regression models have been proposed. Many of these are estimators of so called semiparametric models. One argument for using estimators of semiparametric models is that their application does not require a full parametric specification of the distribution. This property implies a smaller risk of invalid inference due to distributional misspecification compared to, for example, ML estimators.

As heard from the name, semiparametric models combine a parametric part of the model with a nonparametric part so that the model is semiparametric (or semi-nonparametric). There are several definitions of semiparametric models (see e.g., Powell, 1994; Lee, 1996). One common definition, reported by Powell (1994), is that semiparametric models have a finite-dimensional parameter of interest (the parametric part) and an infinite-dimensional nuisance parameter (the nonparametric part).

In semiparametric regression models usually the "deterministic" part of the relationship between the response and explanatory variables have a parametric form, e.g., $X_i^T\beta$ as in (1), while the distribution of the error terms are not specified up to a finite-dimensional vector of parameters. There might be situations where the definition of the term semiparametric models given above is not exact enough, but for the use of the term semiparametric in this thesis it should suffice.

Strictly speaking it is the models that are semiparametric and not the estimators. It might be possible to classify some estimators both as parametric, semiparametric, or nonparametric whereby the use of the term semiparametric estimators perhaps should be avoided.

For example, the OLS estimator of a linear regression model with complete data can be viewed as a "parametric estimator" if the error terms are assumed normally distributed because then the OLS estimator is identical to the ML estimator. On the other hand, the OLS can be viewed as a "semiparametric estimator" if only a zero conditional mean of the error term, i.e., $E(\varepsilon_i|X_i) = 0$, and no parametric distribution of the errors is assumed. That is to say it is the assumptions made about the model that determine whether the estimator is semiparametric or not. Despite this fact, many authors use the term semiparametric estimators and it is also used in this thesis. Here the term semiparametric estimator is defined as an estimator that does not require the error distribution of the model to be known up to a finite-dimensional vector of parameters. The regression function is however specified up to a finite-dimensional vector of parameters, β_0.

Powell (1994) gives a survey on estimation of semiparametric models. Most of the semiparametric models can be characterised by the restriction imposed on the infinite-dimensional component, usually the error distribution. Powell (1994) lists common restrictions and describes in which situations they are useful. The exemplified restrictions are a constant conditional mean, a constant conditional quantile, a constant conditional location, a conditional symmetry, and an independence restriction. Some of these restrictions are also considered and exemplified in Lee (1996).

2.4 Semiparametric estimators

A review of estimators proposed for truncated regression models is found in Lee and Kim (1998). They also present results from a simulation study on the properties of the estimators. They find that the symmetrically trimmed least squares estimator (STLS) (Powell, 1986), the cosine estimator (COS) (Lee and Kim, 1998), and

the quadratic mode estimator (QME) (Lee, 1993) all perform well. These three estimators are all semiparametric and are derived under symmetry assumptions placed on the distribution of the error term. However, Laitila (2001) and Newey (2001) showed that the QME of the slope parameters in a truncated regression model are consistent and asymptotically normal under asymmetrically distributed errors as well. This is a major advantage of the QME. However, a disadvantage of the QME is the complexity in the estimation of the associated covariance matrix, which depends on the density of the errors (Lee, 1993; Lee and Kim, 1998).

Honoré and Powell (1994) present results from a simulation study where estimators suggested for semiparametric censored regression models are compared in terms of bias, mean square error (MSE), and median absolute deviation. They find that the censored least absolute deviation estimator (CLAD) (Powell, 1984), the symmetrically censored least squares estimator (SCLS) (Powell, 1986), and the identically censored least absolute deviations estimator (ICLAD) (Honoré and Powell, 1994) perform best. Another estimator, not included in the comparison, is the winsorized mean estimator (WME) (Lee, 1992). The WME estimator includes CLAD as a special case and is related to the SCLS too. One of the advantages of the WME over the CLAD is that the asymptotic covariance matrix of the WME is easier to estimate because it imposes more assumptions.

There are a number of different non-parameteric estimators suggested for LTRC data (e.g., Gross and Lai, 1996; Park, 2004). Semiparametric estimators are however not so frequently encountered in the literature. One example is Kim and Lai (2000).

Because of the generality of the regularity conditions placed on the distribution of the error term it is difficult to obtain finite sample results for the sampling distributions of the semiparametric estimators. Therefore the inference about the parameters in the regression model is based on large sample theory.

Newey (2001) gives conditions for when a conditional moment restriction in truncated and censored regression models yields consistent and asymptotic normally distributed estimators. He also shows that many of the proposed semiparametric estimators for truncated and censored regression models are (or can be) derived through a conditional moment restriction

$$E\left(m(Y - X^T\beta_0)|X\right) = E\left(m(\varepsilon)|X\right) = 0, \tag{7}$$

where $m(\cdot)$ is a known scalar function. The conditional moment restriction (7) is regarded as the first order condition to a minimisation problem, defining the estimator as the minimum of the corresponding objective function obtained by "integrating back from" $m(\varepsilon)$. For instance, with

$$m(\varepsilon) = 1[-c \leq \varepsilon \leq c]\varepsilon \tag{8}$$

for the left truncated regression model at $t_i = 0$, the QME,

$$\begin{aligned} \hat{\beta}_{QME} &= \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} 1\left[-c < Y_i - \max\left(X_i^T\beta, c\right) < c\right] \\ &\times \left(\left\{Y_i - \max\left(X_i^T\beta, c\right)\right\}^2 - c^2\right), \end{aligned} \tag{9}$$

is obtained. A similar example is

$$m(\varepsilon) = 1[-c \leq \varepsilon \leq c]\varepsilon + 1[\varepsilon > c]c - 1[\varepsilon < -c]c \tag{10}$$

used by Lee (1992) to define the WME,

$$\begin{aligned} \hat{\beta}_{WME} &= \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} 1\left[\left|Y_i - \max\left(X_i^T\beta, c\right)\right| < c\right] \\ &\times \left(0.5(Y_i - \max\left(X_i^T\beta, c\right))^2\right) \\ &+ 1\left[\left|Y_i - \max\left(X_i^T\beta, c\right)\right| \geq c\right] \\ &\times \left(c\left|Y_i - \max\left(X_i^T\beta, c\right)\right| - 0.5c^2\right) \end{aligned} \tag{11}$$

for *left* censoring at $s_i = 0$.

Due to the complexity of the semiparametric estimators and the generality of model regularity conditions, it is difficult to derive finite sample properties of the estimators. Instead, the asymptotic properties are focussed upon and finite sample properties are studied by means of simulation and empirical applications. Properties to consider are, for example, consistency, asymptotic normality of the estimator, $\hat{\beta}$, and consistent estimation of the asymptotic covariance matrix of $\hat{\beta}$. As noted above, finding an estimator of the asymptotic covariance matrix can be difficult for some estimators, such as with the QME. After achieving this, if the sample is large enough, it is possible to make inference about the parameters based on asymptotic results. However, it is also necessary to study how inference based on asymptotic properties of an estimator works in small samples. Good asymptotic properties do not guarantee good finite sample properties. Samples in practice are usually finite.

3 Summary of papers

This thesis contributes in several ways to the existing knowledge on estimation of truncated, censored, and LTRC regression models. Three new semiparametric estimators are proposed, allowing for asymmetric error distributions. Moreover, the problem with estimation of the asymptotic covariance matrix of the QME (9) is considered. In addition, finite sample properties of estimators for truncated, censored, and LTRC data are studied within simulation studies and applications using real data.

The first paper consists of a simulation study of bias and MSE of the QME and other alternative estimators of regression models under truncation. The estimation of the covariance matrix has been considered as a disadvantage of the QME compared to other al-

ternative estimators (Lee, 1993; Lee and Kim, 1998). The paper contributes with results which suggest that the bootstrap technique is potentially useful for estimation of the QME covariance matrix.

In the second paper, estimators of left truncated and right censored semiparametric regression models are proposed. These estimators are generalizations of the QME (9) and the WME (11) by allowing asymmetric "trimming" of observations. Consistency and asymptotic normality of the estimators is shown. Good small sample properties are shown in a simulation study.

By combining the two moment restrictions used to derive the estimators in the second paper, an estimator for left truncated and right censored semiparametric regression models is proposed in the third paper. Consistency of the estimator is shown.

The fourth paper is based on an application where interpurchase intervals of cars is analysed. The data is collected within a panel study of car owners and the interpurchase intervals are left truncated and right censored. The degree of censoring is very high, almost 80 percent. Results regarding consumers' interpurchase behaviour as well as estimators' properties are given.

In the following sections the papers are summarised in more detail.

3.1 Paper I: Finite sample properties of the QME

In this paper a simulation study on the finite sample properties of the QME (9) and a bootstrap estimator of its covariance matrix is presented. The simulation study is based on data about truncated travel distances from the Swedish Travel Habit Survey (Statistics Sweden, 1994). Three different error distributions are considered: a normal, a Gumbel, and a gamma distribution.

The QME is studied with respect to bias and MSE under both symmetric and asymmetric distributions. The ML estimator based

on normality as well as the OLS, STLS, and COS estimators are included for comparison.

As expected from the asymptotic properties, the QME performs better than the STLS and COS estimators under an asymmetric distribution of the errors. The results also indicate that the ML estimator is sensitive to misspecification of the error distribution.

Estimation of the QME covariance matrix is difficult, since it depends on the density of the errors (Lee, 1993; Lee and Kim, 1998). The bootstrap technique has been suggested as an alternative for covariance matrix estimation in regression contexts (e.g., Wu, 1986; Buchinsky, 1995). In this paper, the bootstrap technique is evaluated as a solution to the problem of estimating the covariance matrix of the QME. The bootstrap covariance matrix estimator is defined as

$$\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\beta}_b^{boot}-\bar{\hat{\beta}}^{boot}\right)\left(\hat{\beta}_b^{boot}-\bar{\hat{\beta}}^{boot}\right)^T, \tag{12}$$

where $\bar{\hat{\beta}}^{boot}=\frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b^{boot}$ and $\hat{\beta}_b^{boot}$ are bootstrap replicates computed for each of the B bootstrap samples of n observations (Y_i, X_i^T) sampled with replacement from the original sample. The bootstrap covariance matrix estimator (12) is evaluated in terms of estimated size and power of associated t-tests. Bootstrap replicates are also used for construction of confidence intervals, and their coverage percentages are estimated.

Based on the simulation results it seems as if bootstrapping can provide a solution to the problem of estimating the QME covariance matrix. The evaluated bootstrap techniques work well and overestimate rather than underestimate the variances. t-tests based on the bootstrapped covariance matrix have power, and sizes lower than the nominal size. The basic percentile method for bootstrapping confidence intervals yields coverage higher than the nominal coverage.

The implication of the paper is that the results on bias and MSE of the QME bring with them incitements to find a practical estimator of the covariance matrix of the QME. The bootstrap technique seems to provide such an estimator. However, further research on the properties of the method along with new developments is desirable.

3.2 Paper II: Estimators of regression parameters for truncated and censored data

In this paper two estimators of parameters in semiparametric left truncated and right censored regression models are proposed. In contrast to the majority of existing estimators, the proposed estimators do not require the error term of the regression model to have a symmetric distribution. Therefore they are of more general applicability than many earlier estimators proposed.

The estimators proposed in this paper are called LT and RC and are generalizations of the QME (9) and the WME (11), respectively. The idea is that by allowing asymmetric "trimming" of observations these estimators should be more flexible and more efficient than the QME and the WME because more observations can contribute with possible information to the estimator instead of being trimmed. Both the QME and the WME were first derived under the assumption of symmetry and the definitions of the estimators amounting to "symmetric trimming" of observations is due to this assumption. The idea of using a symmetric "window" $\pm c$ when defining the QME and the WME is that a symmetric window together with a unimodal and symmetric density, for the error term, implies the moment conditions $E[m(\varepsilon)|X] = 0$ defined in (8) and (10) hold. However, the QME and the WME are both consistent for the slope parameters under asymmetrically distributed errors as well (Laitila, 2001; Newey, 2001). Thus, the use of a symmetric window $\pm c$ does not seem to be necessary for defining consistent estimators.

The LT and RC estimators proposed in this paper are derived from the following conditional moment restrictions,

$$E[m(\varepsilon)|X] = E[1[-c_L \leq \varepsilon \leq c_U]\varepsilon|X] = 0 \tag{13}$$

for the truncated regression model and

$$E\left[1[-c_L < \varepsilon < c_U]\varepsilon + 1[\varepsilon \geq c_U]c_U - 1[\varepsilon \leq -c_L]c_L|X\right] = 0 \tag{14}$$

for the censored regression model, where c_L and c_U are positive constants chosen by the researcher. The functions (8) and (10) used to obtain the QME and the WME are special cases of (13) and (14) with $c_L = c_U = c$.

Consistency and asymptotic normality of the LT and RC estimators are shown using the results presented by Newey (2001). Finite sample properties of the estimators are illustrated within a small simulation study and by an empirical application modelling travel distance. The estimators behave well in finite samples with respect to bias and MSE. The results also indicate that the LT and RC estimators with asymmetric windows have a potential to be more efficient and have smaller bias than the QME and the WME with symmetric windows. The estimated parameters in the empirical application are reasonable on the basis of theory about travel distances. However, the choice of threshold values, c_L and c_U, is important for both the finite and large sample results.

3.3 Paper III: A semiparametric regression estimator under left truncation and right censoring

In this paper, a semi-parametric estimator for linear regression models with LTRC data is proposed. The estimator is obtained by combining the moment restrictions (13) and (14) used to obtain the two estimators proposed in Paper II for estimation of slope parameters

of left truncated and right censored regression models, respectively. Starting from this moment condition

$$E[m(\varepsilon)|X] = E[1[-c_L \leq \varepsilon \leq c_U]\varepsilon + 1[\varepsilon > c_U]c_U|X] = 0,$$

where $c_L > 0$ and $c_U > 0$ are constants chosen by the researcher, an objective function $Q_n(\beta)$ is derived. The LTRC estimator is defined as

$$\hat{\beta}_{LTRC} = \arg\min_{\beta \in B} Q_n(\beta) \tag{15}$$

Consistency of the slope parameters is shown under a set of regularity conditions. The proof is based on Lemmas 2.2 and 2.3 in White (1980) by first showing convergence of the objective function $Q_n(\beta)$ in (15) to its expectation uniformly over the parameter space and then that this expectation attains a unique minimum at the true parameter vector β with an unknown constant added to the intercept. Results on asymptotic distribution are not yet available.

Bias, MSE and distributional shape of the LTRC estimator (15) were studied by means of simulation for a regression model under mild truncation and censoring. The results in the simulation study were promising and confirmed the asymptotic result on consistency. However, the results show that the LTRC estimator is sensitive to the choice of the thresholds, c_L and c_U, in the definition of the estimator (15).

3.4 Paper IV: Analysis of interpurchase intervals of cars using truncated and censored data

This paper contributes with analyses of the relation between interpurchase intervals of cars and demographic and economic characteristics of the car owners. The results show that age of the car owner and the household income are most important for the durability of

keeping a car. The interpurchase intervals decrease as household income increases and younger car owners replace their car earlier than older car owners.

Data was collected within a panel study of households in Sweden. Due to the study design the data on the interpurchase intervals is LTRC. Some sampled households had already replaced their car before the start of the panel study and were not included in the study, and many household had not yet replaced their car when the study was terminated. Moreover there were drop-outs during the panel study also resulting in right censored interpurchase intervals. The degree of censoring was almost 80 percent, which made the estimation of the models challenging.

The accelerated failure time (AFT) model (Kalbfleisch and Prentice, 1980) which is a linear regression model with log of the interpurchase intervals as the response variable was used to analyse the data. Bayus and Mehta (1995) recommend using the Weibull and gamma distributions to model interpurchase intervals, because the hazard of a consumer replacing a durable should increase with the duration. In this paper the generalized gamma distribution, which embeds several other distributions, such as the Weibull, gamma, and lognormal distributions as special cases, was used as a model for the distribution of the interpurchase intervals.

The results imply that the models recommended by Bayus and Mehta (1995) for interpurchase intervals of durable goods might be inappropriate for interpurchase intervals of cars. Thus, replacement behaviour might be different among durable goods. The results imply that the assumption of which distribution to use should be handled with a cautious hand. A more flexible approach such as using a semiparametric model or a more general parametric distribution (e.g., the generalized gamma) as a model can offer a solution to the risk of distributional misspecification.

Here, however, the high proportion of censored observations com-

plicated the estimation of the more flexible approaches. The semiparametric LTRC estimator (15) proposed in Paper III did not work when the AFT model could not be estimated by this estimator. The optimization algorithm did not converge to a minimum. In fact, a global maximum of the ML estimator based on generalized gamma distributed interpurchase intervals was not found either. However, as stated above, it was concluded that the Weibull and gamma distributions did not provide this maximum.

In addition to the AFT model, the Cox proportional hazards model (Cox, 1972) was also used to analyse the relationship between interpurchase intervals and characteristics of car owners. The results from this analysis were, in terms of estimated effects of different explanatory variables, similar to the results from the analysis of the AFT model.

4 Further research

This thesis has also raised new and interesting questions. The results show that the behaviour of the three semiparametric estimators, LT, RC, and LTRC, proposed in Papers II and III depend on the "threshold" values c_L and c_U in their definitions. The properties of the QME and the WME also depend on choice of threshold values. Therefore, research on the relation between the threshold values and the properties of the estimators are desired. Furthermore, some "objective rule" on how to choose threshold values is desirable for making the estimator easier to apply in practice.

Thanks to the results of Paper I, the potential of using the bootstrap technique for estimation of the QME covariance matrix should be studied further. For example, is the bootstrap estimator consistent? How does the bootstrap estimator perform for the generalization of the QME (i.e., the LT estimator) proposed in Paper II?

Moreover, new developments are desirable in modifying the bootstrap estimator to perform even better. This could be, for example, an adjustment to the bias of the QME estimator.

Results regarding asymptotic distribution as well as finite sample properties are required for the LTRC estimator proposed in Paper III before it can be used in practice. Further studies on the asymptotic properties of LTRC estimator could also contribute with less strict assumptions for consistency. In addition to simulation studies, it is desirable to compare the LTRC estimator with other estimators in an application with real data to study its finite sample properties. In Paper IV such a comparison is made, but due to the large censoring the LTRC estimator did not work. One idea is to modify estimators, such as the LTRC estimator or ML estimators, under high degree of censoring by giving the complete (i.e., uncensored) observations a larger weight in the function to minimise or maximise. One idea is to choose the weight function such that the modified estimator and the original estimator are asymptotically equivalent.

Finally, the simulation studies and empirical applications included in Papers I–IV have shown that the optimization algorithms and the starting values employed by them are important for results. Especially for simulation studies and for the computer intensive bootstrap technique it is important that the computation time and the accuracy of the algorithms used are acceptable. In this thesis different non-gradient methods have been used to search for the optimum of the non-linear objective functions to obtain estimates for parameters in truncated, censored, and LTRC regression models. It would be interesting to compare different non-gradient and gradient methods for the models and estimators considered in this thesis in a simulation study.

References

[1] Bayus, L. B., Mehta, R. (1995), A segmentation model for targeted marketing of consumer durables, *Journal of Marketing Research* **XXXII**, 463–469.

[2] Buchinsky, M. (1995), Estimating the asymptotic covariance matrix for quantile regression models. A Monte Carlo study, *Journal of Econometrics* **68**, 303–338.

[3] Cox, D. R. (1972), Regression models and life tables (with discussion), *Journal of the Royal Statistics Society* **B 34**, 187–220.

[4] Davidson, R., MacKinnon, J. G. (1993), *Estimation and Inference in Econometrics*. New York: Oxford University Press.

[5] Ferguson, T. S. (1996), *A course in large sample theory*. New York: Chapman & Hall.

[6] Gross, S. T., Lai, T. L. (1996), Nonparametric estimation and regression analysis with left-truncated and right-censored data, *Journal of the American Statistical Association* **91**, 1166–1180.

[7] Hausman, J. A., Wise, D. A. (1977), Social experimentation, truncated distributions, and efficient estimation, *Econometrica* **45**, 919–938.

[8] Honoré B.E., Powell J.L. (1994), Pairwise difference estimators for censored and truncated regression models, *Journal of Econometrics* **64**, 241–278.

[9] Kalbfleisch, J. D., Prentice, R. L. (1980), *The statistical analysis of failure time data*. New York: Wiley.

[10] Kim, C. K., Lai, T. L. (2000), Efficient score estimation and adaptive M-estimators in censored and truncated regression models, *Statistica Sinica* **10**, 731–749.

[11] Laitila, T. (2001), Properties of the QME under asymmetrically distributed disturbances, *Statistics & Probability Letters* **52**, 347–352.

[12] Lee, M. J. (1992), Winsorized mean estimator for censored regression, *Econometric Theory* **8**, 368–382.

[13] Lee, M. J. (1993), Quadratic mode regression, *Journal of Econometrics* **57**, 1–19.

[14] Lee, M. J. (1996), *Methods of Moments and Semiparametric Econometrics for Limited Dependent Variable Models*. New York: Springer.

[15] Lee, M. J. and Kim, H. (1998), Semiparametric econometric estimation for a truncated regression model: a review with an extension, *Statistica Neerlandica* **52**, 200–225.

[16] Newey, W. K. (2001), Conditional moment restrictions in censored and truncated regression models, *Econometric Theory* **17**, 863–888.

[17] Park, J. (2004), Optimal global rate of convergence in nonparametric regression with left-truncated and right-censored data, *Journal of Multivariate Analysis* **89**, 70-86.

[18] Powell, J. L. (1984), Least absolute deviation estimation for the censored regression model, *Journal of Econometrics* **25**, 303–325.

[19] Powell, J. L. (1986), Symmetrically trimmed least squares estimation for tobit models, *Econometrica* **54**, 1435–1460.

[20] Powell, J. L. (1994), Estimation of semiparametric models. In: Engel RF, McFadden DL (Eds.) *Handbook of Econometrics*, Vol 4, pp 2444–2521. Amsterdam: North-Holland.

[21] Statistics Sweden (SCB) (1994), *Svenskarnas resor 1994. Teknisk rapport*. Stockholm: SCB.

[22] Vijverberg, W. P. M. (1987), Non-normality as distributional misspecification in single-equation limited dependent variable models, *Oxford Bullentin of Economics and Statistics* **49**, 417–430.

[23] White, H. (1980), Nonlinear regression on cross-section data, *Econometrica* **48**, 721–746.

[24] White, H. (2001), *Asymptotic Theory for Econometricans*. Revised Edition. San Diego: Academic Press.

[25] Wu, C. F. J. (1986), Jacknife, bootstrap and other resampling methods in regression analysis (with discussion), *The Annals of Statistics* **14**, 1261–1350.