Journal of Statistical Software

May 2014, Volume 57, Issue 14. http://www.jstatsoft.org/

truncSP: An R Package for Estimation of Semi-Parametric Truncated Linear Regression Models

Maria Karlsson
Umeå University

Anita Lindmark
Umeå University

Abstract

Problems with truncated data occur in many areas, complicating estimation and inference. Regarding linear regression models, the ordinary least squares estimator is inconsistent and biased for these types of data and is therefore unsuitable for use. Alternative estimators, designed for the estimation of truncated regression models, have been developed. This paper presents the R package **truncSP**. The package contains functions for the estimation of semi-parametric truncated linear regression models using three different estimators: the symmetrically trimmed least squares, quadratic mode, and left truncated estimators, all of which have been shown to have good asymptotic and finite sample properties. The package also provides functions for the analysis of the estimated models. Data from the environmental sciences are used to illustrate the functions in the package.

Keywords: truncation, limited dependent variable, semi-parametric estimators, R.

1. Introduction

Consider situations in which, for some reason, data on some particular characteristic are not available if their level is below (or above) a fixed limit, but we still need to estimate a regression model with this variable as the response variable. This problem is common in many areas and could occur, e.g., when the measurement equipment is associated with a detection limit. One example is in the environmental sciences, where trace level concentrations, e.g., in air or water, are typically reported as less than a certain limit of detection, t, rather than as actual values when they lie below t. The reason for this approach is that the concentrations are considered as unknown when they cannot be measured accurately (e.g., Lubin *et al.* 2004 and Helsel 2005). Moreover, in some monitoring studies, values less than t are often not even required to be reported to the authorities and hence might be completely missing from the dataset. If the

measurement value is not recorded at all, the data are left truncated at t. If they are recorded but registered only as "smaller than" t, the data are said to be left censored. Censoring and truncation also occur frequently in the area of astronomy, often because of upper detection limits caused by, e.g., sensitivity limit problems in telescopes (Isobe, Feigelson, and Nelson 1986 and Feigelson and Babu 1998). Other examples of a truncated dependent variable include sampling from a subpopulation, e.g., the insurance claims registered at an insurance company (i.e., those insurance damage sizes judged by the insurance holder to have a value greater than the deductible), e.g., Paulsen, Lunde, and Skaug (2008).

This incompleteness of data requires special estimators of the regression coefficients. Several alternative estimators of truncated regression models have been suggested. Many of these are estimators of so-called semi-parametric models, i.e., regression models with the usual parametric relationship between the response and the explanatory variables while the distribution of the error terms is not specified but only assumed to satisfy mild regularity assumptions. We call such estimators semi-parametric estimators, although it is the models and not the estimators that are semi-parametric. For a review and a comparison of properties of suggested estimators for regression models under truncated data, see Lee and Kim (1998). However, despite the many possible application areas and the good asymptotic and finite sample properties of the proposed semi-parametric estimators, to our knowledge, few (if any) have found their way into statistical software. Not even LIMDEP (Greene 2007), which is promoted for its capability to handle truncated regression (among many other things), provides any estimators other than a maximum likelihood estimator assuming normally distributed errors.

In this paper, we present the package **truncSP** for R (R Core Team 2013). **truncSP** is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=truncSP` and contains three semi-parametric estimators for truncated regression models. These are the symmetrically trimmed least squares (STLS) estimator (Powell 1986), the quadratic mode (QME) estimator (Lee 1993 and Laitila 2001), and the left truncated (LT) estimator (Karlsson 2006). All three estimators use trimming of the conditional density of the error terms. The STLS approach assumes symmetrically distributed error terms, whereas the QME and LT approaches have also been shown to be consistent for estimation of the slope parameters under asymmetrically distributed errors (Laitila 2001 and Karlsson 2006). The STLS and QME estimators are included in the comparison by Lee and Kim (1998) who find them to be among the best alternatives. The LT estimator is compared to the QME estimator in Karlsson (2006) where it is concluded that the LT estimator has better performance than the QME estimator in some situations.

The paper is organized as follows. In Section 2, we introduce the truncated regression model and some of the semi-parametric estimators of the regression coefficients suggested in the literature. In Section 3, the R package **truncSP** is described, and in Section 4, data on air pollution and its relationship to traffic and weather characteristics are used to illustrate the package. The data are available in the package **truncSP** and also through StatLib (`http://lib.stat.cmu.edu/`). The paper ends with a summary and concluding remarks.

2. Semi-parametric models and estimators

The form of a linear regression model is

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

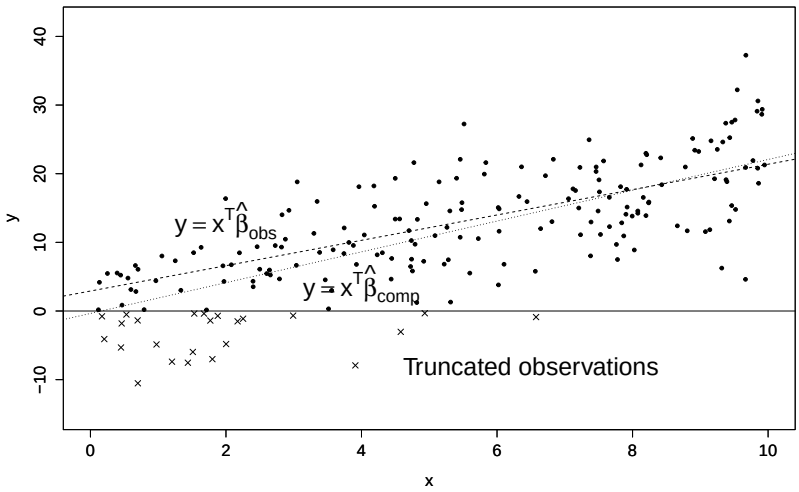


Figure 1: Example of truncated data. OLS estimates using the complete data (dotted line) and the observable data (broken line).

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. The error terms ε_i $(i = 1, \ldots, n)$ are assumed to be independent and identically distributed with mean 0 and variance σ^2.

The objective in 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. Well-known methods for estimating regression models are available in most statistical software packages. The so-called ordinary least squares (OLS) and maximum likelihood (ML) estimators are the most commonly adopted estimators.

Unfortunately, if the response variable is truncated, for reasons such as those mentioned above, this complicates estimation and inference. 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 $Y_i > t$, $(Y_i < t)$ where t is the truncation point. This is equivalent to $\varepsilon_i > t - X_i^\top \beta$ $(\varepsilon_i < t - X_i^\top \beta)$ expressed in error terms. Left truncation and right truncation are sometimes called truncation from below or above, respectively. The regression model (1) with a truncated response variable is known as a truncated regression model. Henceforth, for simplicity, left truncation at $t = 0$ is assumed. The data can be easily transformed into this format if they are right truncated and/or the truncation point is not equal to zero. If the truncation point is $t = a$ $(a \neq 0)$, one subtracts a from the dependent variable to get data with truncation point $t = 0$. To transform right truncated data into left truncated data, one changes the sign of the dependent variable.

Because of the truncation, $\mathsf{E}(\varepsilon|X)$ is not equal to zero but is a function of X; therefore, the OLS estimator is biased and inconsistent and thus not suitable for use. Figure 1 illustrates a simple example with truncated data where the regression line is estimated by OLS using the complete (un-truncated) sample compared to using only the observable data. With a truncated (observable) sample, the OLS underestimates the positive slope of the regression line.

However, an ML estimator, which takes the incompleteness into account, can be used. Consider model (1) under left truncation of Y_i at 0. Let $\mathbf{X}$ denote the $n \times p$ matrix with the

p-dimensional vector of explanatory variables as rows and $\mathbf{y}$ denote the $n \times 1$ vector of the n observations of the truncated response variable. The likelihood function to maximize with respect to (β, η) is then

$$L(\beta, \eta | \mathbf{y}, \mathbf{X}) = \prod_{i=1}^{n} \frac{f(Y_i - X_i^\top \beta | \eta)}{\left(1 - F(-X_i^\top \beta | \eta)\right)}, \tag{2}$$

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 (2) is the density of ε_i conditional on ε_i being included in the sample, i.e., that $\varepsilon_i > -X_i^\top \beta$. However, there are some disadvantages with ML estimation. The PDF $f(\cdot)$ used to formulate the likelihood function has to be chosen. Moreover, for truncated data, the ML estimator is less robust to distributional misspecification than it generally is for complete data (e.g., Davidson and MacKinnon 1993, p. 536). Thus, the choice of which PDF $f(\cdot)$ to use in the expression (2) matters for the properties of the estimator. For truncated regression models, the normal distribution is often used. This is the estimator implemented in the software LIMDEP (Greene 2007) and also in the package **truncreg** (Croissant and Zeileis 2013) in R. Results in Vijverberg (1987) show that assuming a normal distribution in situations when the error distribution is non-normal (focusing on asymmetric non-normality) leads to biased estimates, the bias increasing with the degree of truncation. He recommends the use of alternative estimators under circumstances of non-normality.

2.1. Semi-parametric estimators

Many of the alternative estimators proposed for truncated regression models are so-called semi-parametric estimators. Most of these are derived under the assumption that the error term, ε, in (1) has a symmetric PDF. Because of the truncation of Y (and also of ε), however, we have a sample from a truncated PDF that is, by definition, not symmetric. Symmetry is also the main assumption, together with some other regularity conditions, that is required to establish consistency and asymptotic normality of the semi-parametric estimators STLS and QME (see Theorem 2 in Powell 1986 and Theorem 1 in Lee 1993, respectively), which are two of the estimators available in **truncSP**.

To simplify, the idea is to trim or truncate the observations from above to re-create a symmetric PDF and then basically use the OLS estimator on the remaining observations. Another way to understand the idea behind the estimators is to consider them to be derived through a conditional moment restriction (see Newey 2001, for details),

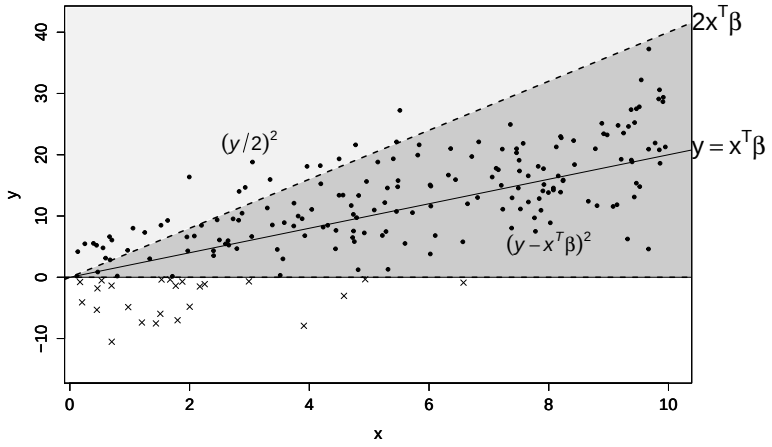
$$\mathsf{E}[m(Y - X^\top \beta_0) | X] = \mathsf{E}[m(\varepsilon) | X] = 0, \tag{3}$$

where $m(\cdot)$ is a known scalar function. The conditional moment restriction is regarded as the first-order condition to a minimization problem that defines the estimator as the minimum of the corresponding objective function, $q(\varepsilon)$, obtained by "integrating back from" $m(\varepsilon)$.

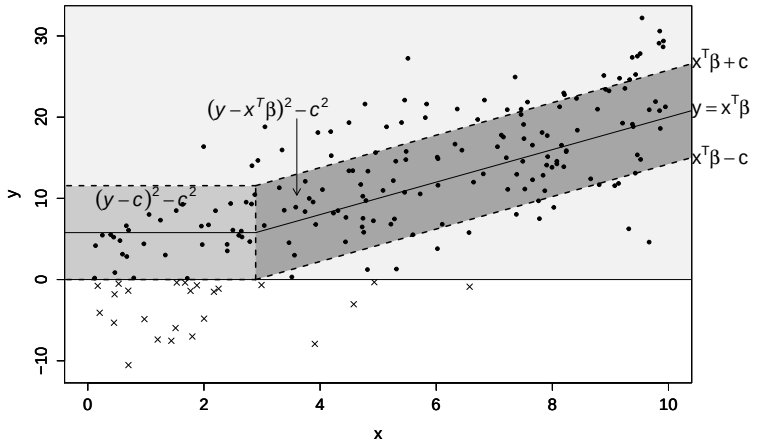
The STLS estimator is defined as

$$\hat{\beta}_{STLS} = \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} q_{STLS}(\varepsilon_i) = \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \max\left(\frac{1}{2} Y_i, X_i^\top \beta \right) \right)^2, \tag{4}$$

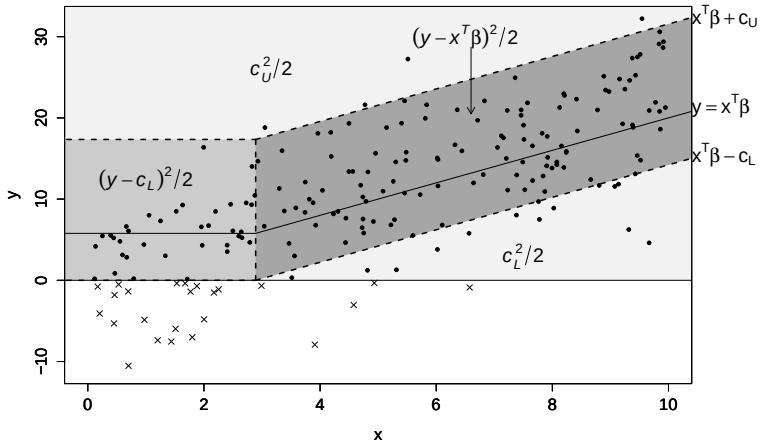
where B denotes the parameter space. This corresponds, in principle, to computing a least squares estimate using only the observations in the darker shaded area in Figure 2(a), i.e.,



(a) STLS



(b) QME



(c) LT

Figure 2: The contributions of the observations (italicized) to the STLS (4), QME (5), and LT (6) objective functions.

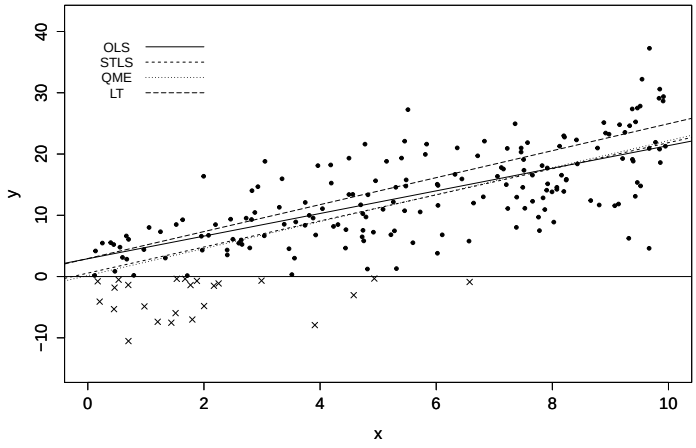


Figure 3: Linear regression functions estimated by OLS, STLS, QME and LT on the truncated data.

trimming the observations where $Y_i > 2X_i^\top \beta$. The observations outside of this area contribute to the objective function through the terms $(Y_i/2)^2$. This means, in practice, that an initial guess of β is required to define the trimming. Then, the objective function is minimized with respect to β to generate an estimate, i.e., a new guess of β, so that a new trimming limit is defined, and so on iteratively until convergence is accomplished. To effectuate this, some type of optimization algorithm must be used. As for most iterative optimization procedures, the initial values can be important for the performance of the optimizer and, in the worst case, the end result. Hence, a "clever" first guess is required to be successful.

Similarly, the QME estimator is defined as

$$\begin{aligned}\hat{\beta}_{QME} &= \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} q_{QME}(\varepsilon_i) \\ &= \arg\min_{\beta \in B} \frac{1}{n} \sum_{i=1}^{n} 1\left[-c < Y_i - \max\left(X_i^\top \beta, c\right) < c\right] \left(\left\{Y_i - \max\left(X_i^\top \beta, c\right)\right\}^2 - c^2\right),\end{aligned} \tag{5}$$

where c is a constant threshold parameter chosen by the researcher and $1[A]$ denotes an indicator function, such that $1[A] = 1$ if condition A holds and 0 otherwise. Figure 2(b) shows the contribution of the observations to the QME objective function. Unlike the STLS approach, the QME approach requires the researcher to choose the amount of trimming, i.e., the value of the threshold c, to use. If c is relatively small, the darkest shaded area in Figure 2(b) is a narrow "belt" containing few observations. If c is relatively large, this area is wide but rather becomes shorter horizontally as the lighter shaded area (containing observations contributing $(y_i - c)^2 - c^2$ to the objective function) becomes longer horizontally. Hence, an intermediate c is preferred. There is no accepted rule for choosing an optimal c. It was suggested in Lee (1993) that it could be desirable to choose c according to some method to minimize, e.g., the MSE. This has not been explored further. In their simulation study, Lee and Kim (1998) used the estimated standard deviation of the observed (i.e., truncated) dependent variable as threshold value with satisfactory results. In a simulation study, Karlsson

(2004) used threshold values based on the estimated (by OLS) residual standard deviation, i.e, the standard deviation of the dependent variable conditional on the explanatory variables.

Laitila (2001) and Newey (2001) showed that the QME estimator of the slope parameters in (1) is also consistent and asymptotically normally distributed under asymmetric distributions of the error term. Karlsson (2006) suggested a version of the QME estimator, which she calls the LT estimator, that trims the observations in a slightly different way to take advantage of the observed data more efficiently than the QME estimator by using an asymmetric trimming window; see Figure 2(c). The LT estimator is defined as

$$\begin{aligned}\hat{\beta}_{LT} &= \arg\min_{\beta\in B} \frac{1}{n}\sum_{i=1}^{n} q_{LT}(\varepsilon_i) \qquad (6)\\ &= \arg\min_{\beta\in B} \sum_{i=1}^{n} 1[X_i^\top\beta > c_L]\cdot\{1[-c_L \le \varepsilon_i \le c_U]\cdot\frac{1}{2}\varepsilon_i^2 + 1[\varepsilon_i < -c_L]\cdot\frac{1}{2}c_L^2 + \\ &\qquad 1[\varepsilon_i > c_U]\cdot\frac{1}{2}c_U^2\} + \\ &\qquad 1[X_i^\top\beta \le c_L]\cdot\{1[-c_L \le Y_i - c_L \le c_U]\cdot\frac{1}{2}(Y_i - c_L)^2 + \\ &\qquad 1[Y_i - c_L < -c_L]\cdot\frac{1}{2}c_L^2 + 1[Y_i - c_L > c_U]\cdot\frac{1}{2}c_U^2\},\end{aligned}$$

where c_L and c_U are the threshold values. The QME estimator can be seen as a special case of the LT estimator with $c_L = c_U = c$, since the scalar functions used to derive the objective function ($m(\cdot)$ in Equation 3) then are identical for both estimators. The choice of symmetric trimming in the QME approach was a direct consequence of assuming a symmetric distribution of the error term when the estimator was first derived. By letting the upper threshold be larger than the lower threshold more observations can possibly contribute their information instead of being trimmed. For further details on how to choose these thresholds see Karlsson (2006). Note that the LT estimator is not consistent for the intercept parameter in the model, nor is the QME under asymmetry.

Lee and Kim (1998) present a review of estimators of truncated regression models, along with a simulation study of their properties. Their results show the STLS estimator, the QME estimator, and their own suggestion (i.e., the cosine (COS) estimator) to be the best estimators. LT is not included in the comparison, but its finite sample behavior is studied and compared to the behavior of QME by means of simulation in Karlsson (2006). The LT estimator is found to behave well in terms of bias and MSE. However, if the error distribution is correctly specified for the ML estimator, e.g., errors are normally distributed and the PDF used in (2) is the normal distribution, then the ML estimator is generally more efficient than these semi-parametric estimators.

Figure 3 shows the linear regression functions estimated by the STLS, QME, and LT approaches. Superimposed is also the OLS estimate, ignoring the truncation, which underestimates the slope. Evident from this figure is that the semi-parametric estimators give more accurate estimates of the slope than the OLS estimator. The LT estimated intercept is off target, as would be expected, but the QME estimated intercept is acceptable because the error term in (1) is symmetric in this example.

2.2. Asymptotic properties

Powell (1986) showed that the STLS estimator is $\sqrt{n}$-consistent and that $\sqrt{n}(\hat{\beta}_{STLS} - \beta_0)$ converges in distribution to $N(0, V_{STLS})$, where V_{STLS} denotes the asymptotic covariance matrix under some regularity conditions, including that the error terms conditionally on X are symmetrically distributed around zero. For details on assumptions and the expression for V_{STLS}, see Theorem 2 in Powell (1986).

Lee (1993) showed that the QME is also $\sqrt{n}$-consistent and that $\sqrt{n}(\hat{\beta}_{QME} - \beta_0)$ converges in distribution to $N(0, V_{QME})$, under symmetrically distributed error terms and some additional assumptions; see Theorem 1 in Lee (1993) for details. However, as noted above, Laitila (2001) showed that the QME is also consistent for β_0 with a unique, but unknown, constant b_c added to the intercept, and asymptotically normally distributed under asymmetrically distributed error terms, under the additional assumption of independence between ε and X. That is, almost surely, $\hat{\beta}_{QME} \to \dot{\beta}_0$ and $\sqrt{n}(\hat{\beta}_{QME} - \dot{\beta}_0) \to N(0, \dot{V}_{QME})$ in distribution, where $\dot{\beta}_0$ is β_0 with b_c added to the intercept. Under a symmetric error distribution, b_c is zero, and $\dot{V}_{QME} = V_{QME}$. Details can be found in Laitila (2001). This finding means that when there is reason to suspect that a symmetric error distribution is not a suitable assumption, the QME estimator still has good properties.

Similarly, Karlsson (2006) showed that the LT estimator is $\sqrt{n}$-consistent for the slope parameters in (1) and that $\sqrt{n}(\hat{\beta}_{LT} - \tilde{\beta}_0)$ converges in distribution to $N(0, V_{LT})$, where $\tilde{\beta}_0$ is the parameter vector β_0 with an unknown constant μ added to the intercept. The LT estimator is also consistent for the intercept plus μ. Note, however, that the intercept can never be consistently estimated by the LT estimator.

The STLS estimator allows for heteroskedasticity of an unknown form, as does the QME estimator if the conditional error term distribution is symmetric. However, the LT and QME estimators, under asymmetrically distributed error terms, do not.

Common to all three estimators is that the expression for their covariance matrices V_{STLS}, V_{QME} (or $\dot{V}_{QME}$), and V_{LT} includes components where the PDF of the error distribution, $f(\cdot)$, occurs. For example,

$$\dot{V}_{QME} = (C_n - D_n)^{-1} A_n (C_n - D_n)^{-1},$$

where

$$\begin{aligned}
A_n &= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left(1[X_i^\top \dot{\beta}_0 - c > 0] 1[-c < Y_i - \dot{\beta}_0 < c](Y_i - X_i\dot{\beta}_0)^2 X_i X_i^\top\right), \\
C_n &= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left(1[X_i^\top \dot{\beta}_0 - c > 0] 1[-c < Y_i - \dot{\beta}_0 < c] X_i X_i^\top\right), \\
D_n &= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}\left(1[X_i^\top \dot{\beta}_0 - c > 0] \left(\frac{c(f(c - b_c) + f(-c + b_c))}{1 - F(-X^\top \beta_0)}\right) X_i X_i^\top\right).
\end{aligned} \tag{7}$$

Because of the occurrence of $f(\cdot)$ in the expression of D_n, estimation of the covariance matrices by substitution of sample moments in place of the expectations in (7) is difficult. An alternative suggested by Lee (1993) and explored by Karlsson (2004) is to use bootstrap techniques. Karlsson (2004) showed that the covariance matrix of the QME estimator can be satisfactorily estimated using the bootstrap. The bootstrap algorithm that was used starts

by taking B bootstrap samples of n observations $\left(Y_i, X_i^\top\right)$ with replacement from the original sample. Next, a bootstrap replicate (i.e., a QME estimate), $\hat{\beta}_b^{boot}$, is calculated for each of the B samples, and finally the covariance matrix is estimated using

$$\frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_b^{boot} - \bar{\hat{\beta}}^{boot})(\hat{\beta}_b^{boot} - \bar{\hat{\beta}}^{boot})^\top, \tag{8}$$

where $\bar{\hat{\beta}}^{boot} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b^{boot}$.

The resulting variances can be used for inference, e.g., significance tests and confidence intervals. Aside from "regular" confidence intervals based on normal approximation, the bootstrap replicates can be used to calculate confidence intervals. A technique for doing this is the so-called percentile method (Efron 1981). To construct a $(1-\alpha)\%$ confidence interval using this method, one simply takes the $(\alpha/2)$th percentile of the bootstrap replicates as the lower limit and the $(1-\alpha/2)$th percentile as the upper limit. Bootstrap confidence intervals with an acceptable coefficient of variation of the estimate usually require more bootstrap replicates than when estimating standard errors using the bootstrap. Based on the recommendations in Efron (1987) $B = 2000$ is used as default in the **truncSP** package.

3. Functions in the package

The package **truncSP** contains three main functions, `stls()`, `qme()` and `lt()`, for estimation using the STLS, QME and LT estimators (see (4), (5) and (6)), respectively.

```
stls(formula, data, point = 0, direction = "left", beta = "ml",
  covar = FALSE, na.action, ...)

qme(formula, data, point = 0, direction = "left", cval = "ml",
  const = 1, beta = "ml", covar = FALSE, na.action, ...)

lt(formula, data, point = 0, direction = "left", clower = "ml",
  const = 1, cupper = 2, beta = "ml", covar = FALSE, na.action, ...)
```

All three functions require an object of class '`formula`', giving a symbolic description of the linear model that is to be estimated (for more information, see the R documentation for class '`formula`').

The arguments `point` and `direction` indicate the truncation point and direction of truncation, respectively. The default values are `point = 0` and `direction = "left"`, respectively. The objective functions used in the optimization were formulated under the assumption that the data are in this form, so if `point` $\neq$ `0` and/or `direction = "right"`, the data are transformed (see Section 2). The resulting estimates are then transformed back to the original scale.

The optimization problems posed by the estimators are solved through the use of the R function `optim()` (from package **stats**, see R Core Team 2013), a general-purpose optimizer. The default, `method = "Nelder-Mead"`, an implementation of the algorithm described in Nelder and Mead (1965), is used. This method has the advantage of also being applicable for

non-differentiable functions, but can be quite slow. The maximum number of iterations is set to `2000`, but the user can alter this through the `...` argument, through which the `control` argument of `optim()` can be adjusted.

The function `optim()` requires starting values for the parameters to be optimized over. These values are provided in the main function call through the argument `beta`. The default method is `"ml"` which uses the estimated coefficients from a truncated maximum likelihood model, assuming Gaussian errors, fitted using the function `truncreg()` (see package **truncreg**, Croissant and Zeileis 2013). Method `"ols"` means that the estimated regression coefficients from fitting a linear regression model through OLS are used. The R function `lm()` (from package **stats**, see R Core Team 2013) is used to provide these. Finally, the starting values can be provided manually as a vector, a column matrix or a row matrix.

The argument `covar` indicates whether or not the covariance matrix of the regression coefficients is estimated. If `TRUE`, the bootstrap (as described in Section 2.2) is used. The functions for this use the function `boot()` from the **boot** package (see Canty and Ripley 2014 and Davison and Hinkley 1997). The number of bootstrap replicates is set to `R = 2000` as the default, but this can be adjusted through the `...` argument, which passes the `R` argument on to `boot()`. The bootstrap procedure can be time-consuming because it requires repeated optimization to produce the bootstrap replicates; therefore, the default is `covar = FALSE`.

Functions `qme()` and `lt()` also have arguments for the threshold values (parameter c in (5) and parameters c_L and c_U in (6)) to be used when trimming the conditional density of the error terms. For `qme()`, the argument `cval` indicates how the threshold value is to be chosen. The methods `"ml"` (the default) and `"ols"` use the estimated residual standard deviation from a truncated regression model fitted with `truncreg()` or from a linear regression model fitted with `lm()`, respectively. Another option is to manually provide the threshold value by supplying a number or numeric vector of length 1. The function `lt()` requires arguments for the upper and lower thresholds. The argument `clower` controls the lower threshold value and has the same options and default value as `cval` for `qme()`. The argument `cupper` determines what upper threshold to use. The user supplies a number that is used to multiply the lower threshold. The default value is `cupper = 2` which means that the upper threshold is two times the size of the lower threshold. Setting `cupper = 1` means that the lower and upper thresholds are the same and that the LT estimates coincide with the QME estimates.

Both `qme()` and `lt()` have an additional argument associated with the threshold values. The argument `const` corresponds to the number by which to multiply `cval` or `clower`. For example, if the user wants to use the QME approach with a threshold value that is half the estimated standard deviation from a model estimated by OLS, he/she supplies the arguments `cval = "ols"`, and `const = 0.5`. Supplying the arguments `clower = "ml"`, `const = 2`, and `cupper = 2` to `lt()` means that the lower threshold will be two times the estimated standard deviation from a truncated maximum likelihood model and that the upper threshold will be two times the size of the lower threshold, i.e., four times the estimated standard deviation. The default value is `const = 1`.

The functions `stls()`, `qme()`, and `lt()` return S4 objects of class '`stls`', '`qme`', and '`lt`', respectively. The objects contain starting values and the estimated coefficients as well as residuals, residual degrees of freedom and fitted values. If `covar = TRUE` they also contain the estimated covariance matrix, the value of `R` used, and the bootstrap replicates. Information from the optimizer, such as the value of the objective function corresponding to the estimated

coefficients and the number of iterations until convergence, is included. Objects of classes '`qme`' and '`lt`' also contain information about the threshold values used.

Methods for analyzing the estimated models are also provided. The extractor functions `coef()`, `residuals()`, `fitted()`, and `vcov()` extract the estimated coefficients, residuals, fitted values, and (if estimated) the covariance matrix of the model. An object from any of the three classes can be summarized through the function

```
summary(object, level = 0.95, ...)
```

If the covariance matrix has been estimated, the summary includes estimated standard deviations and significance tests of the regression coefficients as well as confidence intervals. The confidence intervals are of two different types, the first based on the normal distribution and the second on the percentile method (see Section 2.2). The argument `level` gives the level of confidence for the confidence intervals, with a default of 95%. The function `summary()` returns an S4 object of class '`summary.stls`', '`summary.qme`' or '`summary.lt`', which extends classes '`stls`', '`qme`', and '`lt`', respectively.

4. An illustrative example

This section aims to illustrate the functions in the package as well as to demonstrate differences between the estimators and the effects of using different thresholds values. This is done using data originally collected by the Norwegian Public Roads Administration for a study of air pollution at a road in Oslo, Norway (Aldrin 2006). The dataset `PM10` was initially retrieved from StatLib (`http://lib.stat.cmu.edu/`) and is available in the **truncSP** package. It consists of a subsample of 500 observations from the study and has also been used (among other datasets) as the basis of a simulation study to compare methods of estimating nonlinear functions in additive regression models (Aldrin 2006). The variables in the dataset are

`PM10` hourly values of the logarithm of the concentration of PM_{10} (particles),

`cars` the logarithm of the number of cars per hour,

`temp` the temperature two meters above ground (degrees Celsius),

`wind.speed` wind speed (meters per second),

`temp.diff` the temperature difference between 25 and two meters above ground (degrees Celsius),

`wind.dir` wind direction (between 0 and 360 degrees),

`hour` hour of day,

`day` day number from October 1, 2001.

As previously mentioned, it is common for environmental data to be truncated because of problems in reliably measuring low concentrations. To mimic such a situation and to demonstrate the functions in the package, a new dataset has been generated, eliminating all observations with a `PM10` value of 2 or less. This dataset, `PM10trunc`, contains 460 observations (which corresponds to 8% truncation) and is also available in **truncSP**.

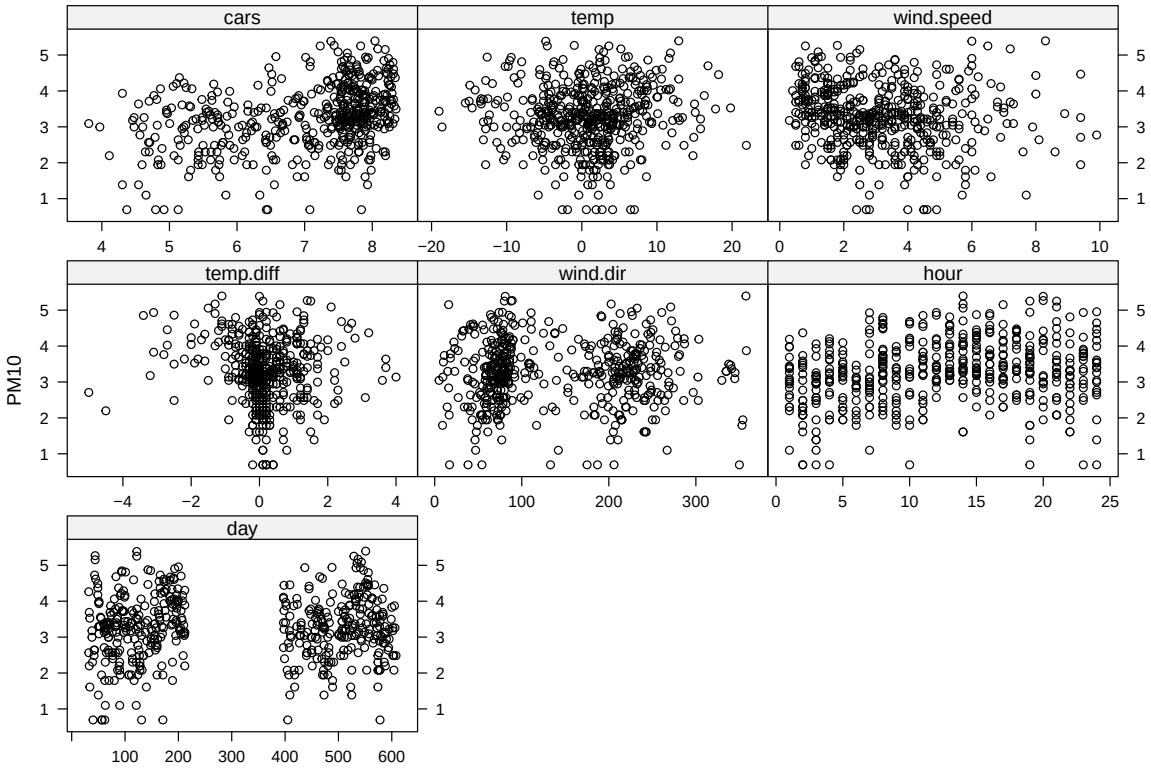


Figure 4: Scatter plots showing the relationships between the variable `PM10` and explanatory variables in the dataset `PM10`.

Figure 4 shows a plot of the explanatory variables in the un-truncated dataset `PM10` against the response variable `PM10`. It is evident that most relationships are nonlinear, indicating that a linear model is not the best solution when modeling these data. However, this analysis is not intended to be used to describe or explain the relationships in the data but rather to illustrate the functions in the package.

4.1. The estimators

To demonstrate the various semi-parametric estimators and compare them to an estimator that is not designed for truncated data, linear models are fitted to the un-truncated dataset using OLS and to the truncated (tr.) dataset using OLS and the three semi-parametric estimators available in **truncSP**. Table 1 shows the estimated regression coefficients from these models as well as the number of iterations (Iter.) required by the semi-parametric methods. The semi-parametric estimates are obtained using the default settings, i.e., estimates from a truncated maximum likelihood model estimated with `truncreg()` are used as the starting values of the vector of regression coefficients, and the estimated residual standard deviation from that same model is used for the threshold / lower threshold in `qme()` and `lt()`. Two times this value is set as the upper threshold in the `lt()` function. It is evident from the results that a large number of iterations is required for the semi-parametric estimators to find the estimates, especially for `qme()`. Concentrating on the estimates of the slope parameters,

	$\hat{\beta}_{\texttt{intercept}}$	$\hat{\beta}_{\texttt{cars}}$	$\hat{\beta}_{\texttt{temp}}$	$\hat{\beta}_{\texttt{wind.speed}}$	$\hat{\beta}_{\texttt{temp.diff}}$	$\hat{\beta}_{\texttt{wind.dir}}$	$\hat{\beta}_{\texttt{hour}}$	$\hat{\beta}_{\texttt{day}}$	Iter.
OLS (full)	1.231*	0.327*	−0.002	−0.103*	0.011	0.000	0.000	0.000	–
OLS (tr.)	2.085*	0.201*	0.001	−0.062¨	−0.004	0.000	0.008	0.000	–
STLS	1.804*	0.247*	0.000	−0.101’	−0.010	0.000	0.009	0.000	1353
QME	1.748*	0.274*	0.002	−0.111¨	0.112	0.000	−0.002	0.000	1885
LT	1.653*	0.307*	0.016	−0.132¨	0.044	0.000	0.011	0.000	1231

Significance codes: p value 0–0.001 = *; 0.001–0.01 = ¨; 0.01–0.05 = ’; 0.05–0.1 = .

Table 1: Estimated coefficients from fitting linear models using different techniques to `PM10` and `PM10trunc`. STLS, QME and LT estimates are derived using the default settings (starting coefficients and threshold values from method `"ml"`, `const = 1` and `cupper = 2`).

	$\hat{\beta}_{\texttt{intercept}}$	$\hat{\beta}_{\texttt{cars}}$	$\hat{\beta}_{\texttt{wind.speed}}$	Iter.
OLS (full)	1.352*	0.321*	−0.105*	–
OLS (tr.)	1.994*	0.230*	−0.064*	–
STLS	1.475’	0.309*	−0.107’	104
QME	2.250¨	0.203’	−0.126¨	178
LT	2.235*	0.246*	−0.104¨	120

Significance codes: p value 0–0.001 = *; 0.001–0.01 = ¨; 0.01–0.05 = ’; 0.05–0.1 = .

Table 2: Estimated coefficients from fitting linear models using different techniques to `PM10` and `PM10trunc`, retaining only the significant variables. STLS, QME and LT estimates are derived using the default settings (starting coefficients and threshold values from method `"ml"`, `const = 1` and `cupper = 2`).

it is clear that none of the models for the truncated data come close to the results from the model of the un-truncated data, although the estimates from the semi-parametric estimators do come closer than the OLS estimate. The variables `cars` and `wind.speed` are significant in all models, although `wind.speed` is less so in the models for the truncated data. Comparing the estimated coefficients for these variables, those from the semi-parametric estimators come closer to the un-truncated model than those from OLS. Overall, it is difficult to say that one estimator gives results that are clearly better than the others for all estimated coefficients using the truncated data.

To simplify comparisons, the analysis is repeated with only the significant variables, yielding the results shown in Table 2. Fewer iterations are required to estimate this smaller model. It is again evident that both variables are significant, although generally less so in the semi-parametric models. It is also clear that the coefficients for `cars` and `wind.speed` from the models estimated using STLS and LT are closer to the values estimated using the full data than the OLS estimates based on the truncated data. QME gives the estimated coefficient for `cars` that is furthest from the OLS estimate based on the full data.

As described in Section 2.1, the threshold values used in `qme()` and `lt()` affect the number of observations used to identify the estimates and thereby the estimates themselves. Table 3 illustrates this, showing the results of different runs of the `qme()` function, all using estimates from a truncated maximum likelihood model as starting values for the vector of regression coefficients but different threshold values. The threshold values have been adjusted using the `const` argument (see Section 3), with `const` ranging from `0.5` to `2`. It is evident that the threshold value chosen has a major impact on the resulting estimates. In this case, setting a

Thresholds	$\hat{\beta}_{\texttt{intercept}}$	$\hat{\beta}_{\texttt{cars}}$	$\hat{\beta}_{\texttt{wind.speed}}$
$0.5 \cdot$ `"ml"`	3.115 (0.772)	0.048 (0.104)	−0.069 (0.049)
$1 \cdot$ `"ml"`	2.250 (0.684)	0.203 (0.089)	−0.126 (0.040)
$2 \cdot$ `"ml"`	1.488 (3.598)	0.320 (0.434)	−0.143 (0.115)

Table 3: Models estimated with the QME approach using different threshold values (standard errors in parentheses).

threshold value at half of the estimated residual standard deviation, and thereby decreasing the number of observations used for identifying the estimates, yields estimates that are far from those in the un-truncated data. Setting two times the standard deviation as the threshold value increases the number of observations used and yields estimates for the intercept and `cars` that are closer to those from the full model than those using `const = 1`. However, using a higher threshold value gives substantially increased standard errors compared to the lower thresholds, and none of the variables are significant.

4.2. Output examples

The basic output from simply running any of the functions `stls()`, `qme()` or `lt()` gives the function call, estimated coefficients, number of iterations, and (for `qme()` and `lt()`) information about the threshold values used.

```
R> library("truncSP")
R> data("PM10trunc", package = "truncSP")
R> qme(formula = PM10 ~ cars + wind.speed, data = PM10trunc, point = 2,
+    covar = TRUE)

Call:
qme(formula = PM10 ~ cars + wind.speed, data = PM10trunc, point = 2,
     covar = TRUE)

Coefficients:
  (Intercept)  cars     wind.speed
   2.2497       0.2030  -0.1256

Iterations:
function
     178

Threshold information:
 Method Constant  Value
     ml        1 0.7599
```

Calling `summary()` on an object from one of the main functions gives output that, at its core, is similar to that from `lm()` or `glm()`, the R functions for fitting linear and generalized linear models (from package **stats**, see R Core Team 2013). Apart from a summary of the estimated regression coefficients, significance tests and confidence intervals are calculated, provided that

the covariance matrix has been estimated. Confidence intervals based on both the normal distribution and the percentile method (see Section 2.2) are provided.

```
R> qmeobj <- qme(formula = PM10 ~ cars + wind.speed, data = PM10trunc,
+    point = 2, covar = TRUE)
R> summary(qmeobj)

Call:
qme(formula = PM10 ~ cars + wind.speed, data = PM10trunc, point = 2,
     covar = TRUE)

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   2.24971     0.68630    3.278   0.00113  **
cars          0.20304     0.08795    2.309   0.02141  *
wind.speed   -0.12564     0.04037   -3.112   0.00197  **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations:
function
     178

Threshold information:
 Method Constant  Value
     ml        1 0.7599


95% Confidence Intervals:
              Lower      Upper
(Intercept)    0.90101    3.59841
cars           0.03021    0.37587
wind.speed    -0.20497   -0.04630

95% Confidence Intervals (Percentile):
              Lower     Upper
(Intercept)    0.4898    2.9970
cars           0.1066    0.4231
wind.speed    -0.2065   -0.0495
```

If the covariance matrix has not been estimated, the output takes the following form.

```
Call:
qme(formula = PM10 ~ cars + wind.speed, data = PM10trunc, point = 2,
     covar = FALSE)
```

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    2.2497          NA       NA        NA
cars           0.2030          NA       NA        NA
wind.speed    -0.1256          NA       NA        NA

Number of iterations:
function
     178

Threshold information:
 Method Constant  Value
     ml        1 0.7599

No covariance matrix has been estimated, hence no t-tests or confidence
intervals are returned. To get these, choose covar=TRUE in the function
call for qme().
```

Finally, a useful extractor function (described briefly in Section 3) available in the package is `vcov()` for extraction of the covariance matrix of the estimated coefficients.

```
R> vcov(qmeobj)

             (Intercept)          cars    wind.speed
(Intercept)  0.471009843 -0.0591754070 -0.0086694016
cars        -0.059175407  0.0077345542  0.0004846089
wind.speed  -0.008669402  0.0004846089  0.0016297807
```

5. Summary and conclusions

Problems with the types of incompleteness of data called truncation occur in many areas and applications. Estimators of so-called semi-parametric truncated regression models have been shown to have good asymptotic and finite sample properties. The practical use of these estimators has been hindered, however, by the lack of software implementations available. Previously, it has fallen to the individual scientist to write code for specific problems with these types of data, but the R package presented in this paper endeavors to provide a more general solution. Three semi-parametric estimators, all of which have been shown to perform well, are included in the package **truncSP**.

Data from the environmental sciences, where applications with truncated data are common, were used to illustrate the functions in the package. Although the data are nonlinear in nature, this served to demonstrate differences between using semi-parametric estimators and using the regular OLS estimator. The effect and importance of carefully choosing the threshold value for trimming of the error terms were also illustrated.

It is our hope that the **truncSP** package will fill a hole in the world of statistical software, as it provides functions for estimation of semi-parametric truncated linear regression models. The

package also has the potential to be developed for nonlinear regression. Karlsson, Cantoni, and de Luna (2009) suggest using local versions of the STLS and QME estimators, inspired by local polynomial regression for un-truncated data (Fan and Gijbels 1996), to estimate nonlinear regressions. The functions in **truncSP** could be adjusted for this purpose by introducing "localizing" weights, as described in Karlsson *et al.* (2009).

References

Aldrin M (2006). "Improved Predictions Penalizing Both Slope and Curvature in Additive Models." *Computational Statistics & Data Analysis*, **50**(2), 267–284.

Canty A, Ripley BD (2014). ***boot**: Bootstrap R (S-PLUS) Functions*. R package version 1.3-11, URL http://CRAN.R-project.org/package=boot.

Croissant Y, Zeileis A (2013). ***truncreg**: Truncated Gaussian Regression Models*. R package version 0.2-1, URL http://CRAN.R-project.org/package=truncreg.

Davidson R, MacKinnon JG (1993). *Estimation and Inference in Econometrics*. Oxford University Press, New York.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge.

Efron B (1981). "Nonparametric Standard Errors and Confidence Intervals." *Canadian Journal of Statistics*, **9**(2), 139–158.

Efron B (1987). "Better Bootstrap Confidence Intervals." *Journal of the American Statistical Association*, **82**(397), 171–185.

Fan J, Gijbels I (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

Feigelson ED, Babu GJ (1998). "Statistical Methodology for Large Astronomical Surveys." In BJ MacLean, DA Golombek, JJE Hayes, HE Payne (eds.), *New Horizons from Multi-Wavelength Sky Surveys. Proceedings of the 179th Symposium of the International Astronomical Union*, pp. 363–370. Kluwer Academic Publisher, Dordrecht.

Greene WH (2007). *LIMDEP 9.0 Reference Guide*. Econometric Software Inc., Plainview, NY. URL http://www.limdep.com/.

Helsel DR (2005). "More Than Obvious: Better Methods for Interpreting Nondetect Data." *Environmental Science and Technology*, **39**(20), 419A–423A.

Isobe T, Feigelson ED, Nelson PI (1986). "Statistical Methods for Astronomical Data with Upper Limits. II – Correlation and Regression." *Astrophysical Journal*, **306**, 490–507.

Karlsson M (2004). "Finite Sample Properties of the QME." *Communication in Statistics – Simulation and Computation*, **33**(3), 567–583.

Karlsson M (2006). "Estimators of Regression Parameters for Truncated and Censored Data." *Metrika*, **63**(3), 329–341.

Karlsson M, Cantoni E, de Luna X (2009). "Local Polynomial Regression with Truncated or Censored Response." *Working Paper 2009:25*, Institute for Labour Market Policy Evaluation (IFAU).

Laitila T (2001). "Properties of the QME Under Asymmetrically Distributed Disturbances." *Statistics & Probability Letters*, **52**(4), 347–352.

Lee MJ (1993). "Quadratic Mode Regression." *Journal of Econometrics*, **57**(1–3), 1–19.

Lee MJ, Kim H (1998). "Semiparametric Econometric Estimators for a Truncated Regression Model: A Review with an Extension." *Statistica Neerlandica*, **52**(2), 200–225.

Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P (2004). "Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits." *Environmental Health Perspectives*, **112**(17), 1691–1996.

Nelder JA, Mead R (1965). "A Simplex Algorithm for Function Minimization." *Computer Journal*, **7**(4), 308–313.

Newey WK (2001). "Conditional Moment Restrictions in Censored and Truncated Regression Models." *Econometric Theory*, **17**(5), 863–888.

Paulsen J, Lunde A, Skaug HJ (2008). "Fitting Mixed-Effects Models When Data are Left Truncated." *Insurance: Mathematics and Economics*, **43**(1), 121–133.

Powell J (1986). "Symmetrically Trimmed Least Squares Estimation for Tobit Models." *Econometrika*, **54**(6), 1435–1460.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Vijverberg WPM (1987). "Non-Normality as Distributional Misspecification in Single-Equation Limited Dependent Variable Models." *Oxford Bulletin of Economics and Statistics*, **49**(4), 417–430.

Affiliation:

Maria Karlsson, Anita Lindmark
Department of Statistics, USBE
Umeå University

SE-901 87 Umeå , Sweden
E-mail: maria.karlsson@stat.umu.se, anita.lindmark@stat.umu.se
URL: http://www.usbe.umu.se/om-handelshogskolan/personal/miakon95
http://www.usbe.umu.se/om-handelshogskolan/personal/anli0053

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 57, Issue 14 *Submitted:* 2011-09-29
May 2014 *Accepted:* 2013-11-22