# Estimating Censored Regression Models in R using the sampleSelection Package

**Arne Henningsen**
University of Copenhagen

**Abstract**

We show how censored regression models (including standard Tobit models) can be estimated in R using the add-on package **sampleSelection**.

*Keywords*: censored regression, Tobit, econometrics, R.

## 1. Introduction

In many statistical analyses of individual data, the dependent variable is censored, e.g. the number of hours worked, the number of extramarital affairs, the number of arrests after release from prison, purchases of durable goods, or expenditures on various commodity groups (Greene 2008, p. 869). If the dependent variable is censored (e.g. zero in the above examples) for a significant fraction of the observations, parameter estimates obtained by conventional regression methods (e.g. OLS) are biased. Consistent estimates can be obtained by the method proposed by Tobin (1958). This approach is usually called "Tobit" model and is a special case of the more general censored regression model.

This paper briefly explains the censored regression model, describes function `tobit` of the R package **sampleSelection**, and demonstrates how this function can be used to estimate censored regression models.

There are also some other functions for estimating censored regression models available in R. For instance function `tobit` from the **AER** package (Kleiber and Zeileis 2008, 2009) and function `cenmle` from the **NADA** package are front ends to the `survreg` function from the **survival** package. Function `tobit` from the **VGAM** package estimates the censored regression model by using its own maximum likelihood routine. Function `MCMCtobit` from the **MCMCpack** package uses the Bayesian Markov Chain Monte Carlo (MCMC) method to estimate censored regression models.

## 2. Censored regression model for cross-sectional data

### 2.1. Standard Tobit model

In the standard Tobit model (Tobin 1958), we have a dependent variable $y$ that is left-censored

at zero:

$$y_i^* = x_i'\beta + \varepsilon_i \tag{1}$$

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ y_i^* & \text{if } y_i^* > 0 \end{cases} \tag{2}$$

Here the subscript $i = 1, \ldots, N$ indicates the observation, $y_i^*$ is an unobserved ("latent") variable, $x_i$ is a vector of explanatory variables, $\beta$ is a vector of unknown parameters, and $\varepsilon_i$ is an error term.

## 2.2. Censored regression model

The censored regression model is a generalisation of the standard Tobit model. The dependent variable can be either left-censored, right-censored, or both left-censored and right-censored, where the lower and/or upper limit of the dependent variable can be any number:

$$y_i^* = x_i'\beta + \varepsilon_i \tag{3}$$

$$y_i = \begin{cases} a & \text{if } y_i^* \leq a \\ y_i^* & \text{if } a < y_i^* < b \\ b & \text{if } y_i^* \geq b \end{cases} \tag{4}$$

Here $a$ is the lower limit and $b$ is the upper limit of the dependent variable. If $a = -\infty$ or $b = \infty$, the dependent variable is not left-censored or right-censored, respectively.

## 2.3. Estimation Method

Censored regression models (including the standard Tobit model) are usually estimated by the Maximum Likelihood (ML) method. Assuming that the error term $\varepsilon$ follows a normal distribution with mean 0 and variance $\sigma^2$, the log-likelihood function is

$$\begin{aligned} \log L = \sum_{i=1}^{N} \Bigg[ & I_i^a \log \Phi\left(\frac{a - x_i'\beta}{\sigma}\right) + I_i^b \log \Phi\left(\frac{x_i'\beta - b}{\sigma}\right) \\ & + \left(1 - I_i^a - I_i^b\right) \left(\log \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right) - \log \sigma\right) \Bigg], \end{aligned} \tag{5}$$

where $\phi(.)$ and $\Phi(.)$ denote the probability density function and the cumulative distribution function, repectively, of the standard normal distribution, and $I_i^a$ and $I_i^b$ are indicator functions with

$$I_i^a = \begin{cases} 1 & \text{if } y_i = a \\ 0 & \text{if } y_i > a \end{cases} \tag{6}$$

$$I_i^b = \begin{cases} 1 & \text{if } y_i = b \\ 0 & \text{if } y_i < b \end{cases} \tag{7}$$

The log-likelihood function of the censored regression model (5) can be maximised with respect to the parameter vector $(\beta', \sigma)'$ using standard non-linear optimisation algorithms.

### 2.4. Implementation in function `tobit`

Censored regression models can be estimated in R with function `tobit`, which is available in the **sampleSelection** package (Toomet and Henningsen 2008). The most important steps done by the `tobit` function are:

1. perform basic checks on the arguments provided by the user
2. prepare the data for the estimation, i.e. the vector of the dependent variable $y = (y_1, \ldots, y_N)'$ and the matrix of the regressors $X = (x_1', \ldots, x_N')'$
3. obtain initial values of the parameters $\beta$ and $\sigma$ from an OLS estimation (if no initial values are provided by the user)
4. define the log-likelihood function (a function that returns the log-likelihood value given the vector of parameters $(\beta', \sigma)'$
5. define a function that calculates the gradients of the log-likelihood function given the vector of parameters[1]
6. call function `maxLik` of the **maxLik** package (Toomet and Henningsen 2010) for the maximisation of the likelihood function
7. add class `"tobit"` to the returned object

### 2.5. Using function `tobit`

Before function `tobit` can be used, the **sampleSelection** package (Toomet and Henningsen 2008) must be loaded:

```
R> library("sampleSelection")
```

The first argument of function `tobit` is `formula`. It is the only mandatory argument and must provide a symbolic description of the model to be fitted. The optional argument `data` can be used to provide a data set (`data.frame`) that contains the variables used in the estimation. We demonstrate the usage of `tobit` by replicating an example given in Kleiber and Zeileis (2008, p. 142). The data used in this example are available in the data set `Affairs` that is included in the R package **AER** (Kleiber and Zeileis 2008, 2009). This data set can be loaded by the following command:

```
R> data("Affairs", package = "AER")
```

In the example of Kleiber and Zeileis (2008, p. 142), the number of a person's extramarital sexual intercourses ("affairs") in the past year is regressed on the person's age, number of years married, religiousness, occupation, and own rating of the marriage. The dependent variable is left-censored at zero and not right-censored. Hence, this is a standard Tobit model. It can be estimated by following command:

[1] The gradients of the log-likelihood function are presented in appendix A.

```
R> tobitResult <- tobit(affairs ~ age + yearsmarried + religiousness +
+     occupation + rating, data = Affairs)
```

Detailed estimation results can be obtained by using the `summary` method.

```
R> summary(tobitResult)
```

```
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 7 iterations
Return code 1: gradient close to zero
Log-Likelihood: -705.5762
7  free parameters
Estimates:
               Estimate Std. error t value   Pr(> t)
(Intercept)    8.174197   2.741446  2.9817  0.002866 **
age           -0.179333   0.079093 -2.2674  0.023368 *
yearsmarried   0.554142   0.134518  4.1195 3.798e-05 ***
religiousness -1.686220   0.403752 -4.1764 2.962e-05 ***
occupation     0.326053   0.254425  1.2815  0.200007
rating        -2.284973   0.407828 -5.6028 2.109e-08 ***
logSigma       2.109859   0.067098 31.4444 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

In case of a censored regression with left-censoring not at zero and/or right-censoring, arguments `left` (defaults to zero) and `right` (defaults to infinity) can be used to specify the limits of the dependent variable. A lower (left) limit of minus infinity (`-Inf`) and an upper (right) limit of infinity (`Inf`) indicate that there is no left-censoring and right-censoring, respectively. For instance, minus the number of extramarital sexual intercourses is not left-censored but right-censored at zero. The same model as above but with the negative number of affairs as the dependent variable can be estimated by

```
R> tobitResultMinus <- tobit(I(-affairs) ~ age + yearsmarried +
+     religiousness + occupation + rating, left = -Inf, right = 0,
+     data = Affairs)
```

This estimation returns $\beta$ parameters that have the opposite sign of the $\beta$ parameters estimated in the original model, but the (logarithmised) standard error of the residuals remains unchanged.

```
R> cbind(coef(tobitResult), coef(tobitResultMinus))
```

```
                   [,1]       [,2]
(Intercept)    8.1741974 -8.1741974
age           -0.1793326  0.1793326
```

```
yearsmarried   0.5541418 -0.5541418
religiousness -1.6862205  1.6862205
occupation     0.3260532 -0.3260532
rating        -2.2849727  2.2849727
logSigma       2.1098592  2.1098592
```

# 3. Censored regression model for panel data

## 3.1. Specification

The censored regression model for panel data with individual specific effects has following specification:

$$y_{it}^* = x_{it}'\beta + \varepsilon_{it} = x_{it}'\beta + \mu_i + \nu_{it} \tag{8}$$

$$y_{it} = \begin{cases} a & \text{if } y_{it}^* \leq a \\ y_{it}^* & \text{if } a < y_{it}^* < b \\ b & \text{if } y_{it}^* \geq b \end{cases} \tag{9}$$

Here the subscript $i = 1, \ldots, N$ indicates the individual, subscript $t = 1, \ldots, T_i$ indicates the time period, $T_i$ is the number of time periods observed for the $i$th individual, $\mu_i$ is a time-invariant individual specific effect, and $\nu_{it}$ is the remaining disturbance.

## 3.2. Fixed effects

In contrast to linear panel data models, we cannot get rid of the individual effects by the within transformation.

## 3.3. Random effects

If the individual specific effects $\mu_i$ are independent of the regressors $x_{it}$, the parameters can be consistently estimated with a random effects model. Assuming that the individual specific effects $\mu$ follow a normal distribution with mean 0 and variance $\sigma_\mu^2$, the remaining disturbance $\nu$ follows a normal distribution with mean 0 and variance $\sigma_\nu^2$, and $\mu$ and $\nu$ are independent, the likelihood contribution of a single individual $i$ is

$$L_i = \int_{-\infty}^{\infty} \left\{ \prod_{t=1}^{T_i} \left[ \Phi\left( \frac{a - x_{it}'\beta - \mu_i}{\sigma_\nu} \right) \right]^{I_{it}^a} \left[ \Phi\left( \frac{x_{it}'\beta + \mu_i - b}{\sigma_\nu} \right) \right]^{I_{it}^b} \left[ \frac{1}{\sigma_\nu} \phi\left( \frac{y_{it} - x_{it}'\beta - \mu_i}{\sigma_\nu} \right) \right]^{\left(1 - I_{it}^a - I_{it}^b\right)} \right\} \phi\left( \frac{\mu_i}{\sigma_\mu} \right) \, d\mu_i \tag{10}$$

and the log-likelihood function is

$$\log L = \sum_{i=1}^{N} \log L_i \tag{11}$$

(see Bruno 2004, p. 2).

Given that we assumed that $\mu$ follows a normal distribution, we can calculate the integrals in the log-likelihood function by the Gauss-Hermite quadrature and then maximise the log-likelihood function using standard non-linear optimisation algorithms (see Butler and Moffit 1982).

Alternatively, the log-likelihood function can be maximized using the method of Maximum Simulated Likelihood (MSL), which allows some flexibility in the specification of the error terms (Greene 2008, p. 799).

*Random effects estimation using the Gauss-Hermite quadrature*

The Gauss-Hermite quadrature is a technique for approximating specific integrals with a weighted sum of function values at some specified points. Applying the Gauss-Hermite quadrature to equation (10), we get

$$L_i = \frac{1}{\sqrt{\pi}} \sum_{h=1}^{H} w_h \left\{ \prod_{t=1}^{T_i} \left[ \Phi \left( \frac{a - x_{it}'\beta - \sqrt{2}\sigma_\mu \psi_h}{\sigma_\nu} \right) \right]^{I_{it}^a} \left[ \Phi \left( \frac{x_{it}'\beta + \sqrt{2}\sigma_\mu \psi_h - b}{\sigma_\nu} \right) \right]^{I_{it}^b} \left[ \frac{1}{\sigma_\nu} \phi \left( \frac{y_{it} - x_{it}'\beta - \sqrt{2}\sigma_\mu \psi_h}{\sigma_\nu} \right) \right]^{\left(1 - I_{it}^a - I_{it}^b\right)} \right\}, \tag{12}$$

where $H$ is number of quadrature points, $\psi_1, \ldots, \psi_H$ are the abscissae, and $w_1, \ldots, w_H$ are the corresponding weights (Greene 2008, p. 553)

# Appendix

## A. Gradients of the log-likelihood function

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{N} \left[ - I_i^a \frac{\phi\left(\frac{a-x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{a-x_i'\beta}{\sigma}\right)} \frac{x_{ij}}{\sigma} + I_i^b \frac{\phi\left(\frac{x_i'\beta-b}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta-b}{\sigma}\right)} \frac{x_{ij}}{\sigma} + \left(1 - I_i^a - I_i^b\right) \frac{y_i - x_i'\beta}{\sigma} \frac{x_{ij}}{\sigma} \right] \tag{13}$$

$$\frac{\partial \log L}{\partial \log \sigma} = \sum_{i=1}^{N} \left[ - I_i^a \frac{\phi\left(\frac{a-x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{a-x_i'\beta}{\sigma}\right)} \frac{a - x_i'\beta}{\sigma} - I_i^b \frac{\phi\left(\frac{x_i'\beta-b}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta-b}{\sigma}\right)} \frac{x_i'\beta - b}{\sigma} + \left(1 - I_i^a - I_i^b\right) \left( \left( \frac{y_i - x_i'\beta}{\sigma} \right)^2 - 1 \right) \right] \tag{14}$$

## References


Bruno G (2004). "Limited Dependent Panel Data Models: A Comparative Analysis of Classical and Bayesian Inference among Econometric Packages." *Computing in Economics and Finance 41*, Society for Computational Economics. URL http://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=SCE2004%&paper_id=41.

Butler J, Moffit R (1982). "A computationally efficient quadrature procedure for the one factor multinomial probit model." *Econometrica*, **50**, 761–764.

Greene WH (2008). *Econometric Analysis*. 6 edition. Prentice Hall.

Kleiber C, Zeileis A (2008). *Applied Econometrics with R*. Springer, New York.

Kleiber C, Zeileis A (2009). *AER: Applied Econometrics with R*. R package version 1.1, http://CRAN.R-project.org/package=AER.

Tobin J (1958). "Estimation of relationship for limited dependent variables." *Econometrica*, **26**, 24–36.

Toomet O, Henningsen A (2008). "Sample Selection Models in R: Package sampleSelection." *Journal of Statistical Software*, **27**(7), 1–23. URL http://www.jstatsoft.org/v27/i07/.

Toomet O, Henningsen A (2010). *maxLik: Tools for Maximum Likelihood Estimation*. R package version 0.7, http://CRAN.R-project.org/package=maxLik.

**Affiliation:**

Arne Henningsen
Institute of Food and Resource Economics
University of Copenhagen
1958 Frederiksberg, Denmark
E-mail: arne.henningsen@gmail.com
URL: http://www.arne-henningsen.name/