Sample Selection Models in R: Package sampleSelection

Ott Toomet
Tartu University

Arne Henningsen
University of Copenhagen

Abstract

This introduction to the R package **sampleSelection** is a slightly modified version of Toomet and Henningsen (2008b), published in the *Journal of Statistical Software*.

This paper describes the implementation of Heckman-type sample selection models in R. We discuss the sample selection problem as well as the Heckman solution to it, and argue that although modern econometrics has non- and semiparametric estimation methods in its toolbox, Heckman models are an integral part of the modern applied analysis and econometrics syllabus. We describe the implementation of these models in the package **sampleSelection** and illustrate the usage of the package on several simulation and real data examples. Our examples demonstrate the effect of exclusion restrictions, identification at infinity and misspecification. We argue that the package can be used both in applied research and teaching.

Keywords: sample selection models, Heckman selection models, econometrics, R.

1. Introduction

Social scientists are often interested in causal effects—what is the impact of a new drug, a certain type of school or being born as a twin. Many of these cases are not under the researcher's control. Often, the subjects can decide themselves, whether they take a drug or which school they attend. They cannot control whether they are twins, but neither can the researcher—the twins may tend to be born in different types of families than singles. All these cases are similar from the statistical point of view. Whatever is the sampling mechanism, from an initial "random" sample we extract a sample of interest, which may not be representative of the population as a whole (see Heckman and MaCurdy 1986, p. 1937, for a discussion).

This problem—people who are "treated" may be different than the rest of the population—is usually referred to as a *sample selection* or *self-selection* problem. We cannot estimate the causal effect, unless we solve the selection problem[1]. Otherwise, we will never know which part of the observable outcome is related to the causal relationship and which part is due to the fact that different people were selected for the treatment and control groups.

Solving sample selection problems requires additional information. This information may be in different forms, each of which may or may not be feasible or useful for any particular case.

[1]Correcting for selectivity is necessary but not always sufficient for estimating the causal effect. Another common problem is the lack of common support between the treated and untreated population. We are grateful to a referee for pointing this out.

Here we list a few popular choices:

- Random experiment, the situation where the participants do not have control over their status but the researcher does. Randomisation is often the best possible method as it is easy to analyse and understand. However, this method is seldom feasible for practical and ethical reasons. Even more, the experimental environment may add additional interference which complicates the analysis.
- Instruments (exclusion restrictions) are in many ways similar to randomisation. These are variables, observable to the researcher, and which determine the treatment status but not the outcome. Unfortunately, these two requirements tend to contradict each other, and only rarely do we have instruments of reasonable quality.
- Information about the functional form of the selection and outcome processes, such as the distribution of the disturbance terms. The original Heckman's solution belongs to this group. However, the functional form assumptions are usually hard to justify.

During recent decades, either randomisation or pseudo-randomisation (natural experiments) have become state of the art for estimating causal effects. However, methods relying on distributional assumptions are still widely used. The reason is obvious—these methods are simple, widely available in software packages, and they are part of the common econometrics syllabus. This is true even though reasonable instruments and parametric assumptions can only seldom be justified, and therefore, it may be hard to disentangle real causal effects from (artificial) effects of parametric assumptions.

Heckman-type selection models also serve as excellent teaching tools. They are extensively explained in many recent econometric text books (e.g. Johnston and DiNardo 1997; Verbeek 2000; Greene 2002; Wooldridge 2003; Cameron and Trivedi 2005) and they are standard procedures in popular software packages like **Limdep** (Greene 2007) and Stata (StataCorp. 2007). These models easily allow us to experiment with selection bias, misspecification, exclusion restrictions etc. They are easy to implement, to visualize, and to understand.

The aim of this paper is to describe the R (R Development Core Team 2008) package **sampleSelection** (version 0.6-0), which is available on the Comprehensive R Archive Network at `https://CRAN.R-project.org/package=sampleSelection`. The package implements two types of more popular Heckman selection models which, as far as we know, were not available for R before. Our presentation is geared toward teaching because we believe that one of the advantages of these types of models lies in econometrics training.

The paper is organized as follows: In the next section we introduce the Heckman (1976) solution to the sample selection problem. Section 3 briefly describes the current implementation of the model in **sampleSelection** and its possible generalisations. In Section 4 we illustrate the usage of the package on various simulated data sets. Section 5 is devoted to replication exercises where we compare our results to examples in the literature. Section 6 describes robustness issues of the method and our implementation of it; and the last section concludes.

2. Heckman's solution

The most popular solutions for sample selection problems are based on Heckman (1976). A variety of generalisations of Heckman's standard sample selection model can be found in the

literature. These models are also called "generalized Tobit models" (Amemiya 1984, 1985). A comprehensive classification of these models has been proposed by Amemiya (1984, 1985).

2.1. Tobit-2 models

Heckman's standard sample selection model is also called "Tobit-2" model (Amemiya 1984, 1985). It consists of the following (unobserved) structural process:

$$y_i^{S*} = \boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S + \varepsilon_i^S \tag{1}$$

$$y_i^{O*} = \boldsymbol{\beta}^{O\prime}\boldsymbol{x}_i^O + \varepsilon_i^O, \tag{2}$$

where y_i^{S*} is the realisation of the the latent value of the selection "tendency" for the individual i, and y_i^{O*} is the latent outcome. $\boldsymbol{x}_i^S$ and $\boldsymbol{x}_i^O$ are explanatory variables for the selection and outcome equation, respectively. $\boldsymbol{x}^S$ and $\boldsymbol{x}^O$ may or may not be equal. We observe

$$y_i^S = \begin{cases} 0 & \text{if } y_i^{S*} < 0 \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

$$y_i^O = \begin{cases} 0 & \text{if } y_i^S = 0 \\ y_i^{O*} & \text{otherwise,} \end{cases} \tag{4}$$

i.e. we observe the outcome only if the latent selection variable y^{S*} is positive. The observed dependence between y^O and x^O can now be written as

$$\mathsf{E}[y^O|\boldsymbol{x}^O = \boldsymbol{x}_i^O, \boldsymbol{x}^S = \boldsymbol{x}_i^S, y^S = 1] = \boldsymbol{\beta}^{O\prime}\boldsymbol{x}_i^O + \mathsf{E}[\varepsilon^O|\varepsilon^S \geq -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S]. \tag{5}$$

Estimating the model above by OLS gives in general biased results, as $\mathsf{E}[\varepsilon^O|\varepsilon^S \geq -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S] \neq 0$, unless ε^O and ε^S are mean independent (in this case $\varrho = 0$ in equation (6) below).

Assuming the error terms follow a bivariate normal distribution:

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^O \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \varrho \\ \varrho & \sigma^2 \end{pmatrix} \right), \tag{6}$$

we may employ the following simple strategy: find the expectations $\mathsf{E}[\varepsilon^O|\varepsilon^S \geq -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S]$, also called the *control function*, by estimating the selection equations (1) and (3) by probit, and thereafter insert these expectations into equation (2) as additional covariates (see Greene 2002 for details). Accordingly, we may write:

$$y_i^O = \boldsymbol{\beta}^{O\prime}\boldsymbol{x}_i^O + \mathsf{E}[\varepsilon^O|\varepsilon^S \geq -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S] + \eta_i \equiv \boldsymbol{\beta}^{O\prime}\boldsymbol{x}_i^O + \varrho\sigma\lambda(\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S) + \eta_i \tag{7}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is commonly referred to as inverse Mill's ratio, $\phi(\cdot)$ and $\Phi(\cdot)$ are standard normal density and cumulative distribution functions and η is a new disturbance term, independent of $\boldsymbol{x}^O$ and $\boldsymbol{x}^S$. The unknown multiplicator $\varrho\sigma$ can be estimated by OLS ($\hat{\beta}^\lambda$). Essentially, we describe the selection problem as an omitted variable problem, with $\lambda(\cdot)$ as the omitted variable. Since the true $\lambda(\cdot)$s in equation (7) are generally unknown, they are replaced by estimated values based on the probit estimation in the first step.

The relations (6) and (7) also reveal the interpretation of ϱ. If $\varrho > 0$, the third term in the right hand side of (7) is positive as the observable observations tend to have above average

realizations of ε^O. This is usually referred to as "positive selection" in a sense that the observed outcomes are "better" than the average. In this case, the OLS estimates are upward biased.

An estimator of the variance of ε^O can be obtained by

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}}}{n^O} + \frac{\sum_i \hat{\delta}_i}{n^O}\,\hat{\beta}^{\lambda\,2} \tag{8}$$

where $\hat{\boldsymbol{\eta}}$ is the vector of residuals from the OLS estimation of (7), n^O is the number of observations in this estimation, and $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \hat{\boldsymbol{\beta}}^{S\prime}\boldsymbol{x}_i^S)$. Finally, an estimator of the correlation between ε^S and ε^O can be obtained by $\hat{\varrho} = \hat{\beta}^{\lambda}/\hat{\sigma}$. Note that $\hat{\varrho}$ can be outside of the $[-1, 1]$ interval.

Since the estimation of (7) is not based on the true but on estimated values of $\lambda(\cdot)$, the standard OLS formula for the coefficient variance-covariance matrix is not appropriate (Heckman 1979, p. 157). A consistent estimate of the variance-covariance matrix can be obtained by

$$\widehat{\mathsf{VAR}}\left[\hat{\boldsymbol{\beta}}^O, \hat{\boldsymbol{\beta}}^{\lambda}\right] = \hat{\sigma}^2 \left[\mathbf{X}_{\lambda}^{O\prime}\mathbf{X}_{\lambda}^{O}\right]^{-1} \left[\mathbf{X}_{\lambda}^{O\prime}\left(\mathbf{I} - \hat{\varrho}^2\hat{\boldsymbol{\Delta}}\right)\mathbf{X}_{\lambda}^{O} + \mathbf{Q}\right] \left[\mathbf{X}_{\lambda}^{O\prime}\mathbf{X}_{\lambda}^{O}\right]^{-1} \tag{9}$$

where

$$\mathbf{Q} = \hat{\varrho}^2 \left(\mathbf{X}_{\lambda}^{O\prime}\hat{\boldsymbol{\Delta}}\mathbf{X}^S\right) \widehat{\mathsf{VAR}}\left[\hat{\boldsymbol{\beta}}^S\right] \left(\mathbf{X}^{S\prime}\hat{\boldsymbol{\Delta}}\mathbf{X}_{\lambda}^{O}\right), \tag{10}$$

$\mathbf{X}^S$ is the matrix of all observations of $\boldsymbol{x}^S$, $\mathbf{X}_{\lambda}^O$ is the matrix of all observations of $\boldsymbol{x}^O$ and $\hat{\lambda}$, $\mathbf{I}$ is an identity matrix, $\hat{\boldsymbol{\Delta}}$ is a diagonal matrix with all $\hat{\delta}_i$ on its diagonal, and $\widehat{\mathsf{VAR}}\left[\hat{\boldsymbol{\beta}}^S\right]$ is the estimated variance covariance matrix of the probit estimate (Greene 1981, 2002).

This is the original idea by Heckman (1976). As the model is fully parametric, it is straightforward to construct a more efficient maximum likelihood (ML) estimator. Using the properties of a bivariate normal distribution, it is easy to show that the log-likelihood can be written as

$$\ell = \sum_{\{i:y_i^S=0\}} \log\Phi(-\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S) + \tag{11}$$

$$+ \sum_{\{i:y_i^S=1\}} \left[\log\Phi\left(\frac{\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S + \frac{\varrho}{\sigma}(y_i^O - \boldsymbol{\beta}^{O\prime}\boldsymbol{x}_i^O)}{\sqrt{1-\varrho^2}}\right) - \frac{1}{2}\log 2\pi - \log\sigma - \frac{1}{2}\frac{(y_i^O - \boldsymbol{\beta}^{O\prime}\boldsymbol{x}_i^O)^2}{\sigma^2}\right]. \tag{12}$$

The original article suggests using the two-step solution for exploratory work and as initial values for ML estimation, since in those days the cost of the two-step solution was \$15 while that of the maximum-likelihood solution was \$700 (Heckman 1976, p. 490). Nowadays, costs are no longer an issue, however, the two-step solution allows certain generalisations more easily than ML, and is more robust in certain circumstances (see Section 6 below).

This model and its derivations were introduced in the 1970s and 1980s. The model is well identified if the exclusion restriction is fulfilled, i.e. if $\boldsymbol{x}^S$ includes a component with a substantial explanatory power but which is not present in $\boldsymbol{x}^O$. This means essentially that we have a valid instrument. If this is not the case, the identification is related to the non-linearity of the inverse Mill's ratio $\lambda(\cdot)$. The exact form of it stems from the distributional assumptions. During the recent decades, various semiparametric estimation techniques have been

increasingly used in addition to the Heckman model (see Powell 1994, Pagan and Ullah 1999, and Li and Racine 2007 for a review).

2.2. Tobit-5 models

A straightforward generalisation of the standard sample selection model (Tobit-2) is the switching regression (Tobit-5) model. In this case, we have two outcome variables, where only one of them is observable, depending on the selection process. Switching regression problems arise in a wide variety of contexts, e.g. in treatment effect, migration or schooling choice analysis. This type of model consists of a system of three simultaneous latent equations:

$$y_i^{S*} = \boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S + \varepsilon_i^S \tag{13}$$

$$y_i^{O1*} = \boldsymbol{\beta}^{O1\prime}\boldsymbol{x}_i^{O1} + \varepsilon_i^{O1} \tag{14}$$

$$y_i^{O2*} = \boldsymbol{\beta}^{O2\prime}\boldsymbol{x}_i^{O2} + \varepsilon_i^{O2}, \tag{15}$$

where y^{S*} is the selection "tendency" as in the case of Tobit-2 models, and y^{O1*} and y^{O2*} are the latent outcomes, only one of which is observable, depending on the sign of y^{S*}. Hence we observe

$$y_i^S = \begin{cases} 0 & \text{if } y_i^{S*} < 0 \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

$$y_i^O = \begin{cases} y_i^{O1*} & \text{if } y_i^S = 0 \\ y_i^{O2*} & \text{otherwise.} \end{cases} \tag{17}$$

Assuming that the disturbance terms have a 3-dimensional normal distribution,

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon^{O1} \\ \varepsilon^{O2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \varrho_1\sigma_1 & \varrho_2\sigma_2 \\ \varrho_1\sigma_1 & \sigma_1^2 & \varrho_{12}\sigma_1\sigma_2 \\ \varrho_2\sigma_2 & \varrho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right), \tag{18}$$

it is straightforward to construct analogous two-step estimators as in (7). We may write

$$\mathsf{E}[y^O|\boldsymbol{x}^{O1} = \boldsymbol{x}_i^{O1}, \boldsymbol{x}^S = \boldsymbol{x}_i^S, y^S = 0] = \boldsymbol{\beta}^{O1\prime}\boldsymbol{x}_i^{O1} + \mathsf{E}[\varepsilon^{O1}|\varepsilon^S < -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S] \tag{19}$$

$$\mathsf{E}[y^O|\boldsymbol{x}^{O2} = \boldsymbol{x}_i^{O2}, \boldsymbol{x}^S = \boldsymbol{x}_i^S, y^S = 1] = \boldsymbol{\beta}^{O2\prime}\boldsymbol{x}_i^{O2} + \mathsf{E}[\varepsilon^{O2}|\varepsilon^S \geq -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S] \tag{20}$$

and hence

$$y_i^O = \begin{cases} \boldsymbol{\beta}^{O1\prime}\boldsymbol{x}_i^{O1} - \varrho_1\sigma_1\lambda(-\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S) + \eta_i^1 & \text{if } y_i^S = 0 \\ \boldsymbol{\beta}^{O2\prime}\boldsymbol{x}_i^{O2} + \varrho_2\sigma_2\lambda(\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S) + \eta_i^2 & \text{otherwise,} \end{cases} \tag{21}$$

where $\mathsf{E}[\eta^1] = \mathsf{E}[\eta^2] = 0$ and $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the inverse Mill's ratio which can be calculated using the probit estimate of (16). This system can be estimated as two independent linear models, one for the case $y^S = 0$ and another for $y^S = 1$.

Note that the inverse Mill's ratio enters (21) with opposite signs. If $\varrho_2 > 0$, we find that those, for whom we observe the outcome 2, have more positive realizations of ε^{O2} in average. As the outcome 1 being observable is in the opposite end of the latent y^{S*} scale, upward bias for y^{O1} occurs when $\varrho_1 < 0$. This is what we expect to observe in case of endogenous selection,

a situation where the individuals try to select the "best" one between two possible options, based on some private information about y^{O1*} and y^{O2*}.

The log-likelihood for this problem can be written as

$$\begin{aligned}\ell = &-\frac{N}{2}\log 2\pi + \\ &+ \sum_{\{i:y_i^S=0\}} \left\{ -\log\sigma_1 - \frac{1}{2}\left(\frac{y_i^O - \boldsymbol{\beta}^{O1\prime}\boldsymbol{x}_i^{O1}}{\sigma_1}\right)^2 + \log\Phi\left[-\frac{\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S + \frac{\varrho_1}{\sigma_1}\left(y_i^O - \boldsymbol{\beta}^{O1\prime}\boldsymbol{x}_i^{O1}\right)}{\sqrt{1-\varrho_1^2}} \right] \right\} \\ &+ \sum_{\{i:y_i^S=1\}} \left\{ -\log\sigma_2 - \frac{1}{2}\left(\frac{y_i^O - \boldsymbol{\beta}^{O2\prime}\boldsymbol{x}_i^{O2}}{\sigma_2}\right)^2 + \log\Phi\left[\frac{\boldsymbol{\beta}^{S\prime}\boldsymbol{x}_i^S + \frac{\varrho_2}{\sigma_2}\left(y_i^O - \boldsymbol{\beta}^{O2\prime}\boldsymbol{x}_i^{O2}\right)}{\sqrt{1-\varrho_2^2}} \right] \right\}\end{aligned} \tag{22}$$

where N is the total number of observations. Note that ϱ_{12} plays no role in this model; the observable distributions are determined by the correlations ϱ_1 and ϱ_2 between the disturbances of the selection equation (ε^S) and the corresponding outcome equation (ε^{O1} and ε^{O2}).

3. Implementation in sampleSelection

3.1. Current implementation

The main frontend for the estimation of selection models in **sampleSelection** is the command `selection`. It requires a formula for the selection equation (argument `selection`), and a formula (or a list of two for switching regression models) for the outcome equation (`outcome`). One can choose the method (`method`) to be either `"ml"` for the ML estimation, or `"2step"` for the two-step method. If the user does not provide initial values (`start`) for the ML estimation, `selection` calculates consistent initial values by the corresponding two-step method.

While the actual two-step estimation is done by the function `heckit2fit` or `heckit5fit` (depending on the model), the ML estimation is done by `tobit2fit` or `tobit5fit`. Note that log-likelihood functions of selection models are in general not globally concave, and hence one should use a good choice of initial values (see the example in Section 6.1).

The likelihood maximisation is performed by the **maxLik** package (Toomet and Henningsen 2008a), where the Newton-Raphson algorithm (implemented as the function `maxNR`) using an analytic score vector and an analytic Hessian matrix is used by default. This results in a reasonably fast computation even in cases of tens of thousands observations. A well-defined model should converge in less than 15 iterations; in the case of weak identification this number may be considerably larger. Convergence issues may appear at the boundary of the parameter space (if $|\varrho| \to 1$, see Section 6.1). Other maximisation algorithms can be chosen by argument `maxMethod`, e.g. `maxMethod="SANN"` uses a variant of the robust "simulated annealing" stochastic global optimization algorithm proposed by Bélisle (1992) and `maxMethod="BHHH"` uses the Berndt-Hall-Hall-Hausman algorithm (Berndt, Hall, Hall, and Hausman 1974).

The command `selection` returns an object of class `selection`. The **sampleSelection** package provides several methods for objects of this class: a `print` method prints the estimation results, `summary` method (and associated `print` method) calculate and print summary results,

`coef` methods extract the estimated coefficients, a `vcov` method extracts their covariance matrix, a `fitted` method extracts the fitted values, a `residuals` method extracts the residuals, a `model.frame` method extracts the model frame, and a `model.matrix` method extracts the model matrix.

The `coef` and `vcov` methods for `selection` objects, as well as the `print` method for `summary.selection` objects include an extra argument `part`, which specifies which part of the model is returned or printed. One may choose either the full model (`part="full"`, default), or the outcome equation only (`part="outcome"`). The `fitted`, `residuals`, and `model.matrix` methods also include a `part` argument. However, for these functions the `part` argument specifies whether the objects of the outcome part (`part="outcome"`, default) or of the selection part (`part="selection"`) should be returned.

Currently, the variance-covariance matrix of the two-step estimators is only partially implemented with `NA`-s in place of the unimplemented components.

The package is written completely in R which should minimize the portability issues. It depends on packages **maxLik** (Toomet and Henningsen 2008a), **systemfit** (Henningsen and Hamann 2007a,b), and **mvtnorm** (Genz, Bretz, and Hothorn 2005), where **mvtnorm** is used for examples only. All these packages are available from the Comprehensive R Archive Network at `https://CRAN.R-project.org/`.

3.2. Current API and a wider range of selection models

We now briefly discuss possible ways of introducing more general selection models using a slightly generalized version of the current API.

First, the current argument `selection` can be used for specifying more general selection conditions. `selection` might contain an interval for interval censoring (for instance `selection = c(0, Inf)` in case of the standard Tobit model), more than one formula (for multiple treatment models), or an indicator for the selection mechanism (like `"max"` or `"min"` for switching regression with unobserved separator). In this way, all generalized Tobit models listed by Amemiya (1984, 1985) can be specified.

Another possible generalisation is allowing for discrete dependent variable models. While the case of binary-choice can be easily distinguished from continuous response, we need an additional parameter in the multiple-choice case. This parameter (possibly a vector where components correspond to the individual equations) will allow the user to specify the exact type of the response (like multinominal, ordered or Poissonian).

Third, different distributions of the disturbance terms can be specified in a similar way using an additional parameter. It may be a vector if different marginal distributions for different outcome equations are necessary.

Finally, as the `selection` currently supports only two easily distinguishable models, we have not provided a way to specify the model explicitly. However, explicit specification would allow users to locate the programming problems more easily, and lessen the complications related to automatic guess of the correct model type.

4. Using the `selection` function

This section provides selected illustrative simulation experiments which illustrate both the

strong and weak sides of the method, and the typical usage of `selection`.

4.1. Tobit-2 models

First, we estimate a correctly specified Tobit-2 model with exclusion restriction:

```
R>   set.seed(0)
R>   library("sampleSelection")
R>   library("mvtnorm")
R>   eps <- rmvnorm(500, c(0,0), matrix(c(1,-0.7,-0.7,1), 2, 2))
R>   xs <- runif(500)
R>   ys <- xs + eps[,1] > 0
R>   xo <- runif(500)
R> yoX <- xo + eps[,2]
R> yo <- yoX*(ys > 0)
```

We use **mvtnorm** in order to create bivariate normal disturbances with correlation -0.7. Next, we generate a uniformly distributed explanatory variable for the selection equation, `xs`, the selection outcome `ys` by probit data generating process, and the explanatory variable for the outcome equation `xo`. All our true intercepts are equal to 0 and our true slopes are equal to 1, both in this and the following examples. We retain the latent outcome variable (`yoX`) for the figure below, and calculate the observable outcome `yo`. Note that the vectors of explanatory variables for the selection (`xs`) and outcome equation (`xo`) are independent and hence the exclusion restriction is fulfilled. This can also be seen from the fact that the estimates are reasonably precise:

```
R>   summary( selection(ys~xs, yo ~xo))
```

```
--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 5 iterations
Return code 1: gradient close to zero (gradtol)
Log-Likelihood: -712.3163
500 observations (172 censored and 328 observed)
6 free parameters (df = 494)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2228     0.1081  -2.061   0.0399 *
xs            1.3377     0.2014   6.642 8.18e-11 ***
Outcome equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0002265  0.1294178  -0.002    0.999
xo           0.7299070  0.1635925   4.462 1.01e-05 ***
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma   0.9190     0.0574  16.009  < 2e-16 ***
```

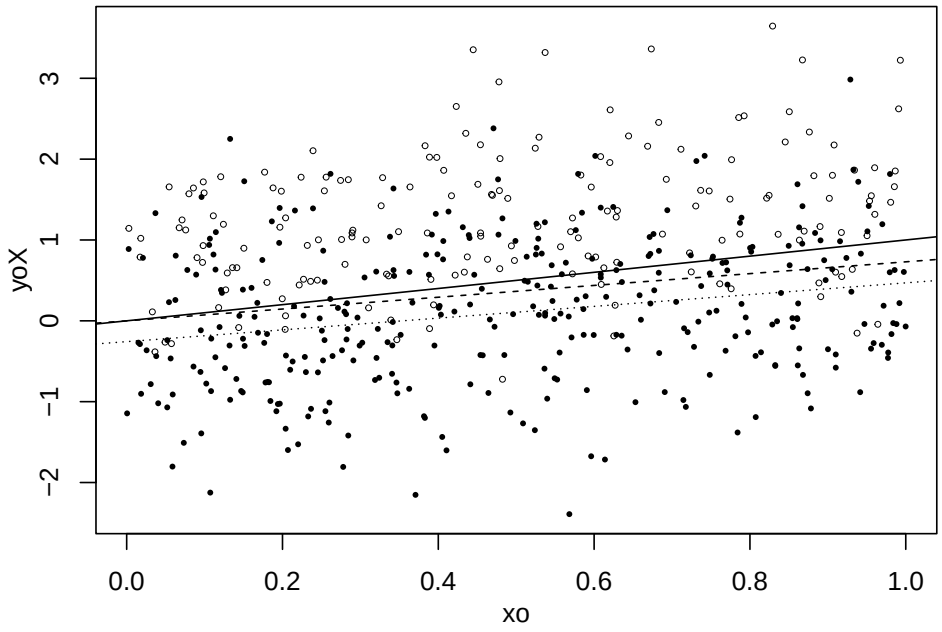


Figure 1: Sample selection example with exclusion restriction (filled circles = observable outcomes, empty circles = unobservable outcomes, solid line = true dependence, dashed line (partly overlapping the solid) = ML estimate above, dotted line = OLS estimate based on observable outcomes only).

```
rho    -0.5392     0.1521  -3.544 0.000431 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

One can see that all the true values are within the 95% confidence intervals of the corresponding estimates.

Now look at the graphical representation of the data (Figure 1). We can see that the unobserved values (empty circles) tend to have higher y^{O*} realisations than the observed ones (filled circles). This is because $\varrho < 0$. The OLS estimate (dotted line) is substantially downward biased – it does not take into account the fact, that we tend to observe the observations with low realisations of ε^O. The slope, however, remains unbiased because $\mathsf{E}[\varepsilon^O | \varepsilon^S \geq -\boldsymbol{\beta}^{S\prime} \boldsymbol{x}_i^S]$ does not depend on $\boldsymbol{x}^O$.

Now we repeat the same exercise, but without the exclusion restriction, generating `yo` using `xs` instead of `xo`.

```
R> yoX <- xs + eps[,2]
R> yo <- yoX*(ys > 0)
R> summary(selection(ys ~ xs, yo ~ xs))

--------------------------------------------
Tobit 2 model (sample selection model)
```

```
Maximum Likelihood estimation
Newton-Raphson maximisation, 14 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -712.8298
500 observations (172 censored and 328 observed)
6 free parameters (df = 494)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1984     0.1114  -1.781   0.0756 .
xs            1.2907     0.2085   6.191 1.25e-09 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5499     0.5644  -0.974  0.33038
xs            1.3987     0.4482   3.120  0.00191 **
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  0.85091    0.05352  15.899   <2e-16 ***
rho   -0.13226    0.72684  -0.182    0.856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

The estimates are still unbiased but standard errors are substantially larger in this case. The exclusion restriction—independent information about the selection process—has a certain identifying power that we now have lost. We are solely relying on the functional form identification.

We illustrate this case with an analogous figure (Figure 2). The selection model uncovers the true dependence very well. The OLS estimate is downward biased because of $\varrho < 0$, as in the previous case. However, now the slope is upward biased because $\mathsf{E}[\varepsilon^O | \varepsilon^S \geq -\boldsymbol{\beta}^{S\prime} \boldsymbol{x}_i^S]$ is increasing in the single explanatory variable in the outcome model, $\boldsymbol{x}^S$.

In order to identify β^λ and $\boldsymbol{\beta}^{O\prime}$ without the exclusion restriction, $\lambda(\cdot)$ must differ from a linear combination of $\boldsymbol{x}^O$ components (see Leung and Yu 1996). The degree of non-linearity in $\lambda(\cdot)$ depends on the variability in $\boldsymbol{\beta}^{S\prime} \boldsymbol{x}^S$ as $\lambda(\cdot)$ is a smooth convex function (see Figure 4). Hence the standard errors of the estimates depend on the variation in the latent selection equation (1), even without the exclusion restriction fulfilled. More variation gives smaller standard errors[2]. We demonstrate this below: Change the support of $\boldsymbol{x}^S$ from $[0, 1]$ to $[-5, 5]$:

```
R>   xs <- runif(500, -5, 5)
R>   ys <- xs + eps[,1] > 0
R> yoX <- xs + eps[,2]
R>   yo <- yoX*(ys > 0)
R>   summary( selection(ys ~ xs, yo ~ xs))
```

[2]The exact shape of the function $\lambda(\cdot)$ is dependent on the distribution of the disturbances. However, $\mathsf{E}[\varepsilon^O | \varepsilon^S \geq -\boldsymbol{\beta}^{S\prime} \boldsymbol{x}^S] \to 0$ as $\boldsymbol{\beta}^{S\prime} \boldsymbol{x}^S \to \infty$ under a wide set of distribution functions. Hence the estimator is less dependent on functional form assumptions if the variability in the latent selection equation is larger. This is related to identification at infinity (Chamberlain 1986).

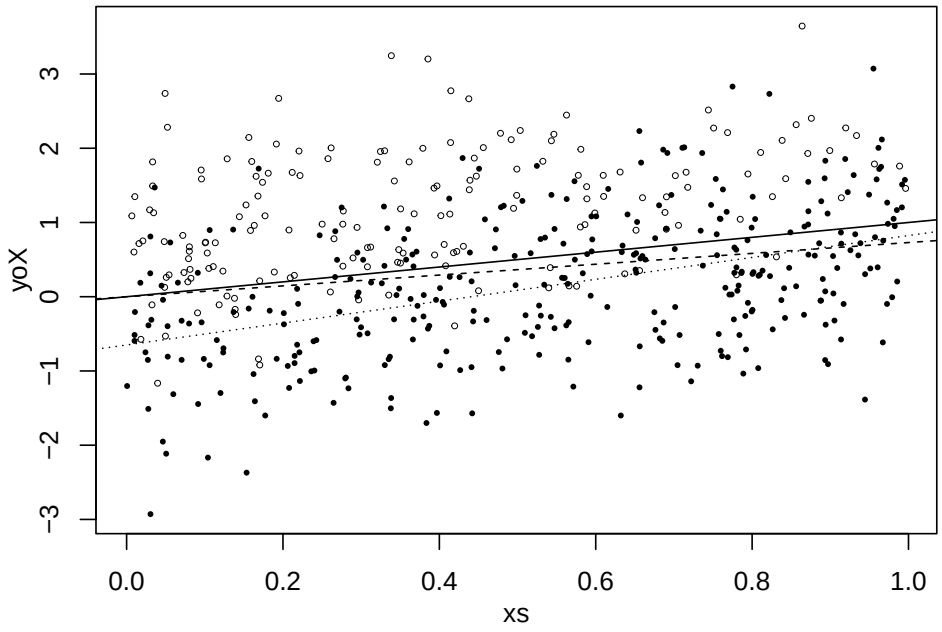


Figure 2: Sample selection example without exclusion restriction (for symbols see Figure 1).

```
--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 4 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -440.883
500 observations (247 censored and 253 observed)
6 free parameters (df = 494)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07401    0.10527  -0.703    0.482
xs           0.98825    0.08860  11.154   <2e-16 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19333    0.14462   1.337    0.182
xs           0.94646    0.04575  20.689   <2e-16 ***
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  1.04038    0.04928  21.110  < 2e-16 ***
rho   -0.77495    0.09433  -8.216 1.87e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

Now all the parameters are precisely estimated, with even higher precision than in the first

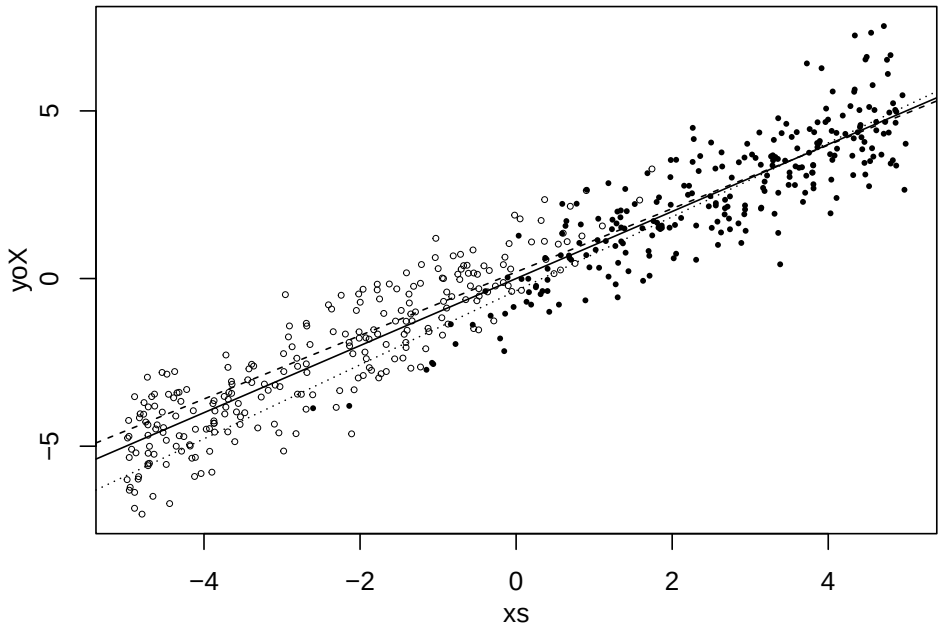


Figure 3: Sample selection example with more variation in x^S (for symbols see Figure 1).

example where the exclusion restriction is fulfilled. The reason is simple: As one can see from Figure 3, selection is not an issue if $x^S > 2$ while virtually nothing is observed if $x^S < -2$. Here $\lambda(\beta^S x^S)$ differs enough from a linear function.

4.2. Switching regression models

Now let us focus on the Tobit-5 examples. We create the following simple switching regression problem:

```
R>   set.seed(0)
R>   vc <- diag(3)
R>   vc[lower.tri(vc)] <- c(0.9, 0.5, 0.1)
R>   vc[upper.tri(vc)] <- vc[lower.tri(vc)]
R>   eps <- rmvnorm(500, c(0,0,0), vc)
R>   xs <- runif(500)
R>   ys <- xs + eps[,1] > 0
R>   xo1 <- runif(500)
R>   yo1 <- xo1 + eps[,2]
R>   xo2 <- runif(500)
R>   yo2 <- xo2 + eps[,3]
```

We generate 3 disturbance vectors by a 3-dimensional normal distribution using `rmvnorm`. We set the correlation between ε^S and ε^{O1} equal to 0.9 and between ε^S and ε^{O2} to 0.5. The third correlation, 0.1, takes care of the positive definiteness of the covariance matrix and does not affect the results. Further, we create three independent explanatory variables (`xs`, `xo1` and `xo2`, uniformly distributed on $[0, 1]$), and hence the exclusion restriction is fulfilled.

`selection` now expects three formulas, one for the selection equation, as before, and a list of two for the outcome equation. Note that we do not have to equalize the unobserved values to zero, those are simply ignored. The results look as follows:

```
R>   summary(selection(ys~xs, list(yo1 ~ xo1, yo2 ~ xo2)))

--------------------------------------------
Tobit 5 model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 11 iterations
Return code 1: gradient close to zero (gradtol)
Log-Likelihood: -895.8201
500 observations: 172 selection 1 (FALSE) and 328 selection 2 (TRUE)
10 free parameters (df = 490)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1550     0.1051  -1.474    0.141
xs            1.1408     0.1785   6.390 3.86e-10 ***
Outcome equation 1:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.02708    0.16395   0.165    0.869
xo1          0.83959    0.14968   5.609  3.4e-08 ***
Outcome equation 2:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1583     0.1885   0.840    0.401
xo2           0.8375     0.1707   4.908 1.26e-06 ***
   Error terms:
       Estimate Std. Error t value Pr(>|t|)
sigma1  0.93191    0.09211  10.118   <2e-16 ***
sigma2  0.90697    0.04434  20.455   <2e-16 ***
rho1    0.88988    0.05353  16.623   <2e-16 ***
rho2    0.17695    0.33139   0.534    0.594
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

We can see that the parameters are fairly well estimated. All the estimates are close to the true values.

Next, take an example of functional form misspecification. We create the disturbances as 3-variate χ^2_1 random variables (we subtract 1 in order to get the mean zero disturbances), and generate `xs` to be in the interval $[-1, 0]$ in order to get an asymmetric distribution over observed choices:

```
R> set.seed(5)
R> eps <- rmvnorm(1000, rep(0, 3), vc)
R> eps <- eps^2 - 1
R> xs <- runif(1000, -1, 0)
```

```
R> ys <- xs + eps[,1] > 0
R> xo1 <- runif(1000)
R> yo1 <- xo1 + eps[,2]
R> xo2 <- runif(1000)
R> yo2 <- xo2 + eps[,3]
R> summary(selection(ys~xs, list(yo1 ~ xo1, yo2 ~ xo2), iterlim=20))

--------------------------------------------
Tobit 5 model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 4 iterations
Return code 3: Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.
Log-Likelihood: -1665.936
1000 observations: 760 selection 1 (FALSE) and 240 selection 2 (TRUE)
10 free parameters (df = 990)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.53698    0.05808  -9.245  < 2e-16 ***
xs           0.31268    0.09395   3.328 0.000906 ***
Outcome equation 1:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.70679    0.03573  -19.78   <2e-16 ***
xo1          0.91603    0.05626   16.28   <2e-16 ***
Outcome equation 2:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1446        NaN     NaN      NaN
xo2           1.1196     0.5014   2.233   0.0258 *
   Error terms:
       Estimate Std. Error t value Pr(>|t|)
sigma1  0.67770    0.01760   38.50   <2e-16 ***
sigma2  2.31432    0.07615   30.39   <2e-16 ***
rho1   -0.97137        NaN     NaN      NaN
rho2    0.17039        NaN     NaN      NaN
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

Although we still have an exclusion restriction, now serious problems appear—most intercepts are statistically significantly different from the true values zero. This model has serious convergence problems and often it does not converge at all (this is why we increased the `iterlim` and used `set.seed(5)`).

As the last Tobit example, we repeat the previous exercise without the exclusion restriction, and a slightly larger variance of `xs`:

```
R> set.seed(6)
```

```
R> xs <- runif(1000, -1, 1)
R>   ys <- xs + eps[,1] > 0
R>   yo1 <- xs + eps[,2]
R>   yo2 <- xs + eps[,3]
R> summary(tmp <- selection(ys~xs, list(yo1 ~ xs, yo2 ~ xs), iterlim=20))

--------------------------------------------
Tobit 5 model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 16 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -1936.431
1000 observations: 626 selection 1 (FALSE) and 374 selection 2 (TRUE)
10 free parameters (df = 990)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3528     0.0424  -8.321 2.86e-16 ***
xs            0.8354     0.0756  11.050  < 2e-16 ***
Outcome equation 1:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.55448    0.06339  -8.748   <2e-16 ***
xs           0.81764    0.06048  13.519   <2e-16 ***
Outcome equation 2:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.6457     0.4994   1.293    0.196
xs            0.3520     0.3197   1.101    0.271
   Error terms:
       Estimate Std. Error t value Pr(>|t|)
sigma1  0.59187    0.01853  31.935   <2e-16 ***
sigma2  1.97257    0.07228  27.289   <2e-16 ***
rho1    0.15568    0.15914   0.978    0.328
rho2   -0.01541    0.23370  -0.066    0.947
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

In most cases, this model does not converge. However, if it does (like in this case, where we use `set.seed(6)`), the results may be seriously biased. Note that the first outcome parameters have low standard errors, but a substantial bias. We present the graph of the correct control function, based on the χ^2 distribution, and one where we assume the normal distribution of the disturbance terms in Figure 4. We use the estimated parameters for constructing the latter, however, we scale the normal control functions (inverse Mill's ratios) to a roughly similar scale as the correct ones.

One can see that the functions differ substantially in the relevant range of $x^S \in [-1, 1]$. In particular, the true $\mathsf{E}[\varepsilon^{O2}|\varepsilon^S \geq -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}^S]$ decreases substantially faster close to $x^S = 1$ than the normal approximation while the correct $\mathsf{E}[\varepsilon^{O1}|\varepsilon^S < -\boldsymbol{\beta}^{S\prime}\boldsymbol{x}^S]$ is decreasing slower compared to the approximation.

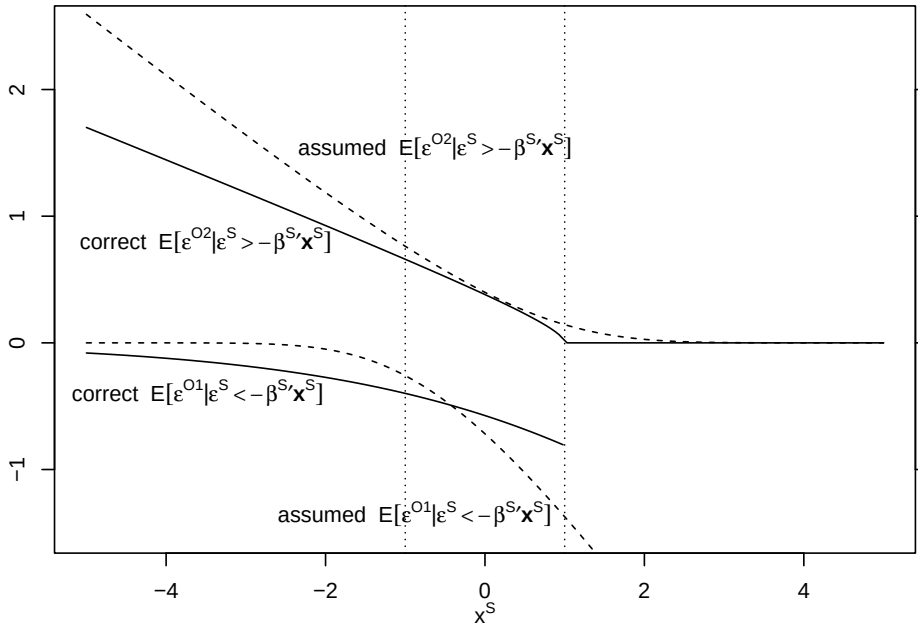


Figure 4: Correct and assumed control functions. Dotted vertical lines denote the support of x^S in the simulation experiment; correct control functions are based on the $\chi^2(1)$ distribution; assumed control functions are based on the normal distribution.

It is instructive to estimate the same model as two independent OLS equations:

```
R>   coef(summary(lm(yo1~xs, subset=ys==0)))

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.6109093 0.02462008 -24.81346 4.339319e-95
xs           0.7773989 0.04403325  17.65482 7.007157e-57

R>   coef(summary(lm(yo2~xs, subset=ys==1)))

             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 0.6136357  0.1130488 5.428058 1.029707e-07
xs          0.3693103  0.1834578 2.013054 4.482896e-02
```

One can see that the OLS estimates are very close to the ML ones. This is related to the fact that none of the ϱs is statistically significantly different from zero.

5. Two replication exercises

In this section we test the reliability of the results from `selection` by applying the two-step and the ML estimation method re-estimating selected models already published in the literature.

5.1. Greene (2002): Example 22.8, page 786

The first test is example 22.8 from Greene (2002, p. 786). The data set used in this example is included in **sampleSelection**; it is called `Mroz87`. This data set was used by Mroz (1987) for analysing female labour supply. In this example, labour force participation (described by dummy `lfp`) is modelled by a quadratic polynomial in age (`age`), family income (`faminc`, in 1975 dollars), presence of children (`kids`), and education in years (`educ`). The wage equation includes a quadratic polynomial in experience (`exper`), education in years (`educ`), and residence in a big city (`city`). First, we have to create a dummy variable for presence of children.

```
R> data( "Mroz87" )
R> Mroz87$kids <- ( Mroz87$kids5 + Mroz87$kids618 > 0 )
```

Now, we estimate the model by the two-step method.

```
R> greeneTS <- selection( lfp ~ age + I( age^2 ) + faminc + kids + educ,
+    wage ~ exper + I( exper^2 ) + educ + city,
+    data = Mroz87, method = "2step" )
```

Most results are identical to the values reported by Greene (2002, p. 786). Only the coefficient of the inverse Mill's ratio ($\beta^{\lambda} = \varrho\sigma$), its standard error, and ϱ deviate from the published results, but all differences are less than one percent.[3]

Finally, we repeat the analysis with the ML estimation method:

```
R> greeneML <- selection( lfp ~ age + I( age^2 ) + faminc + kids + educ,
+    wage ~ exper + I( exper^2 ) + educ + city, data = Mroz87,
+    maxMethod = "BHHH", iterlim = 500 )
```

Again, the estimated coefficients and standard errors are almost identical to the values published in Greene (2006). While we can obtain the same coefficients with the Newton-Raphson (NR) maximisation method, we have to use the Berndt-Hall-Hall-Hausman (BHHH) method to obtain the published standard errors[4]. This is because different ways of calculating the Hessian matrix may result in substantially different standard errors (Calzolari and Fiorentini 1993). The NR algorithm uses exact analytic Hessian, BHHH uses outer product approximation.

5.2. Cameron and Trivedi (2005): Section 16.6, page 553

The data set used in this example is based on the "RAND Health Insurance Experiment" (Newhouse 1999). It is included in **sampleSelection**, where it is called `RandHIE`. Cameron and Trivedi (2005, p. 553) use these data to analyse health expenditures. The endogenous variable of the outcome equation measures the log of the medical expenses of the individual (`lnmeddol`) and the endogenous variable of the selection equation indicates whether the medical expenses are positive (`binexp`). The regressors are the log of the coinsurance rate

[3] Note that the standard error of the coefficient of the inverse Mill's ratio ($\beta^{\lambda} = \varrho\sigma$) is wrong in Greene (2002, p. 786) (see Greene 2006).

[4] We are grateful to William Greene for pointing this out.

plus 1 (`logc = log(coins+1)`), a dummy for individual deductible plans (`idp`), the log of the participation incentive payment (`lpi`), an artificial variable (`fmde` that is 0 if `idp` = 1 and `ln(max(1,mde/(0.01*coins)))` otherwise (where `mde` is the maximum expenditure offer), physical limitations (`physlm`), the number of chronic diseases (`disea`), dummy variables for good (`hlthg`), fair (`hlthf`), and poor (`hlthp`) self-rated health (where the baseline is excellent self-rated health), the log of family income (`linc`), the log of family size (`lfam`), education of household head in years (`educdec`), age of the individual in years (`xage`), a dummy variable for female individuals (`female`), a dummy variable for individuals younger than 18 years (`child`), a dummy variable for female individuals younger than 18 years (`fchild`), and a dummy variable for black household heads (`black`). First, we select the subsample (study year `year` equal to 2 and education information present) that is used by Cameron and Trivedi (2005) and specify the selection as well as the outcome equation.

```
R> data( "RandHIE" )
R> subsample <- RandHIE$year == 2 & !is.na( RandHIE$educdec )
R> selectEq <- binexp ~ logc + idp + lpi + fmde + physlm + disea +
+    hlthg + hlthf + hlthp + linc + lfam + educdec + xage + female +
+    child + fchild + black
R> outcomeEq <- lnmeddol ~ logc + idp + lpi + fmde + physlm + disea +
+    hlthg + hlthf + hlthp + linc + lfam + educdec + xage + female +
+    child + fchild + black
```

Now, we estimate the model by the two-step method (reporting only the coefficients):

```
R> rhieTS <- selection( selectEq, outcomeEq, data = RandHIE[ subsample, ],
+                           method = "2step" )
```

All coefficients and standard errors are fully identical to the results reported by Cameron and Trivedi (2005) — even if they are compared with the seven-digit values in their Stata output that is available on `https://cameron.econ.ucdavis.edu/mmabook/mma16p3selection.txt`.[5]

Again, we repeat the analysis with the ML estimation method:

```
R> rhieML <- selection( selectEq, outcomeEq, data = RandHIE[ subsample, ] )
```

All coefficients and standard errors of the ML estimation are nearly identical to the values reported in Table 16.1 of Cameron and Trivedi (2005) as well as to the seven-digit values in their Stata output. Only a few coefficients deviate at the seventh decimal place.

6. Robustness issues

6.1. Convergence

The log-likelihood function of the models above is not globally concave. The model may not converge, or it may converge to a local maximum, if the initial values are not chosen well

[5]The coefficient and t-value of `idp` in column `lnmed` of Cameron and Trivedi's Table 16.1 seem to be wrong, because they differ from their Stata output as well as from our results.

enough. This may easily happen as we illustrate below. Recall example 22.8 from Greene (2002, p. 786) (section: 5.1). This model gives reasonable results, but these are sensitive to the start values. Now we re-estimate the model specifying start values "by hand" (note that you have to supply a positive initial value for the variance):

```
R> greeneStart <- selection( lfp ~ age + I( age^2 ) + faminc + kids + educ,
+    wage ~ exper + I( exper^2 ) + educ + city,
+    data = Mroz87, start = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.5, 0.9))
R> cat( greeneStart$message )

Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.

R> coef( summary( greeneStart ) )[ "rho", ]

  Estimate Std. Error    t value   Pr(>|t|)
 0.9999997        NaN        NaN        NaN
```

The process did not converge. In the current case the problem lies with the numerical problems at the boundary of the parameter space (note that ϱ is close to 1). A workaround is to use a more robust maximisation method. For instance, one may choose to run the SANN maximizer for 10 000 iterations, and then use the returned coefficients as start values for the Newton-Raphson algorithm.[6]

```
R> set.seed(0)
R> greeneSANN <- selection( lfp ~ age + I( age^2 ) + faminc + kids + educ,
+    wage ~ exper + I( exper^2 ) + educ + city,
+    data = Mroz87, start = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.5, 0.9),
+    maxMethod="SANN", parscale = 0.001 )
R> greeneStartSANN <- selection( lfp ~ age + I( age^2 ) + faminc + kids + educ,
+    wage ~ exper + I( exper^2 ) + educ + city,
+    data = Mroz87, start = coef( greeneSANN ) )
R> cat( greeneStartSANN$message )

successive function values within tolerance limit (tol)
```

The new Newton-Raphson estimate converged to another maximum with a log-likehood value even higher than the one of the original estimate published in Greene (2002, p. 786) (see Section 5.1):

```
R> logLik( greeneML )

'log Lik.' -1581.258 (df=13)

R> logLik( greeneStartSANN )
```

[6] One has to choose a suitable value for `parscale`; `parscale=0.001` worked well for this example.

```
'log Lik.' -1479.654 (df=13)
```

However, in most cases the 2-step method does a good job in calculating initial values.

6.2. Boundary of the parameter space

In general, one should prefer `method="ml"` instead of `"2step"`. However, ML estimators may have problems at the boundary of the parameter space. Take the textbook Tobit example:

$$y_i^* = \boldsymbol{\beta}'\boldsymbol{x}_i + \varepsilon_i; \quad y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

This model can be written as a Tobit-2 model where the error term of the selection and outcome equation are perfectly correlated. In this case the ML estimator may not converge:

```
R> set.seed(0)
R> x <- runif(1000)
R> y <- x + rnorm(1000)
R> ys <- y > 0
R> tobitML <- selection(ys~x, y~x)
R> cat( tobitML$message )

Last step could not find a value above the current.
Boundary of parameter space?
Consider switching to a more robust optimisation method temporarily.

R> coef( summary( tobitML ) )[ "rho", ]

  Estimate Std. Error    t value   Pr(>|t|)
 0.9999994        NaN        NaN        NaN
```

The reason, as in the previous example, is that $\varrho = 1$ lies at the boundary of the parameter space. However, the 2-step method still works, although standard errors are large and $\rho \notin [-1, 1]$:

```
R> tobitTS <- selection(ys~x, y~x, method="2step")
R> coef( summary( tobitTS ) )[ "rho", ]

  Estimate Std. Error    t value   Pr(>|t|)
  1.150939         NA         NA         NA
```

7. Conclusions

This paper describes Heckman-type selection models and their implementation in the package **sampleSelection** for the programming language R. These models are popular in estimating

impacts of various factors in economics and other social sciences. We argue that they also serve as useful teaching tools because they are easy to implement and understand.

We describe the implementation and usage of standard sample selection (Tobit-2) and switching regression (Tobit-5) models in **sampleSelection**, and possible generalisations of our `selection` function. We demonstrate the usage of the package using a number of simulated and real data examples. The examples illustrate several important issues related to exclusion restrictions, identification at infinity, and functional form specification. Our implementation works well for correctly specified cases with the exclusion restriction fulfilled. The problems appearing in the case of misspecification or weak identification are related to the model itself. In these problematic cases, the user may gain from a more robust maximisation algorithm. In some cases, the two-step estimator is preferable.

Acknowledgments

The authors thank Roger Koenker, Achim Zeileis, and two anonymous referees for helpful comments and suggestions. Ott Toomet is grateful to the project TMJRI 0525 2003-2007 of the Estonian Ministry of Education and Science.

References

Amemiya T (1984). "Tobit Models: A Survey." *Journal of Econometrics*, **24**, 3–61.

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

Bélisle CJP (1992). "Convergence Theorems for a Class of Simulated Annealing Algorithms on R^d." *Journal of Applied Probability*, **29**, 885–895.

Berndt EK, Hall BH, Hall RE, Hausman JA (1974). "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement*, **3**(4), 653–665.

Calzolari G, Fiorentini G (1993). "Alternative Covariance Estimators of the Standard Tobit Model." *Economics Letters*, **42**(1), 5–13.

Cameron AC, Trivedi PK (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Chamberlain G (1986). "Asymptotic Efficiency in Semi-Parametric Models with Censoring." *Journal of Econometrics*, **32**, 189–218.

Genz A, Bretz F, Hothorn T (2005). ***mvtnorm**: Multivariate Normal and t Distribution*. R package version 0.7-6, URL https://CRAN.R-project.org/package=mvtnorm.

Greene WH (1981). "Sample Selection Bias as a Specification Error: A Comment." *Econometrica*, **49**(3), 795–798.

Greene WH (2002). *Econometric Analysis*. 5th edition. Prentice Hall. ISBN 0130661899.

Greene WH (2006). "Errata and Discussion of Econometric Analysis, 5th edition." https://pages.stern.nyu.edu/~wgreene/Text/Errata/ERRATA5.htm.

Greene WH (2007). ***Limdep*** *9.0 Econometric Modeling Guide, Volume 1*. Econometric Software, Inc., Plainview, NY. URL https://en.wikipedia.org/wiki/LIMDEP.

Heckman JJ (1976). "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement*, **5**(4), 475–492.

Heckman JJ (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, **47**(1), 153–161.

Heckman JJ, MaCurdy TE (1986). "Labor Econometrics." In Z Griliches, MD Intriligator (eds.), *Handbook of Econometrics*, volume 3, chapter 32, pp. 1917–1977. Elsevier, Amsterdam.

Henningsen A, Hamann JD (2007a). "**systemfit**: A Package for Estimating Systems of Simultaneous Equations in R." *Journal of Statistical Software*, **23**(4), 1–40. URL https://www.jstatsoft.org/v23/i04/.

Henningsen A, Hamann JD (2007b). ***systemfit****: Simultaneous Equation Estimation*. R package version 1.0, URL https://CRAN.R-project.org/package=systemfit.

Johnston J, DiNardo J (1997). *Econometric Methods*. 4th edition. McGraw-Hill.

Leung SF, Yu S (1996). "On the Choice Between Sample Selection and Two-Part Models." *Journal of Econometrics*, **72**, 197–229.

Li Q, Racine JS (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, Princeton.

Mroz TA (1987). "The Sensitivity of an Empirical Model of Married Women's Hours to Work to Economic and Statistical Assumptions." *Econometrica*, **55**(4), 765–799.

Newhouse JP (1999). "RAND Health Insurance Experiment [in Metropolitan and Non-Metropolitan Areas of the United States], 1974–1982." *Aggregated Claims Series, Volume 1: Codebook for Fee-for-Service Annual Expenditures and Visit Counts ICPSR 6439*, ICPSR Inter-university Consortium for Political and Social Research.

Pagan A, Ullah A (1999). *Nonparametric Econometrics*. Themes in Modern Econometrics. Cambridge University Press, Cambridge.

Powell JL (1994). "Estimation of Semiparametric Models." In RF Engle, DL McFadden (eds.), *Handbook of Econometrics*, volume 4, chapter 41, pp. 2443–2521. Elsevier.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL https://www.R-project.org/.

StataCorp (2007). *Stata Statistical Software: Release 10*. StataCorp LP, College Station, Texas. URL https://www.Stata.com/.

Toomet O, Henningsen A (2008a). ***maxLik**: Maximum Likelihood Estimation*. R package version 0.5, URL https://CRAN.R-project.org/package=maxLik.

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

Verbeek M (2000). *A Guide to Modern Econometrics*. John Wiley & Sons, Hoboken, NJ.

Wooldridge JM (2003). *Introductory Econometrics. A Modern Approach*. 2nd edition. Thomson South-Western.

Affiliation:

Ott Toomet
Department of Economics
Tartu University
Narva 4-A123
Tartu 51009, Estonia
Telephone: +372/737.6348
E-mail: otoomet@ut.ee
URL: https://www.obs.ee/~siim/

Arne Henningsen
Institute of Food and Resource Economics
University of Copenhagen
Rolighedsvej 25
1958 Frederiksberg C, Denmark
E-mail: arne.henningsen@gmail.com
URL: https://www.arne-henningsen.name/