Treatment Effects with Normal Disturbances in `sampleSelection` Package

Ott Toomet
University of Washington

October 6, 2026

1 The Problem

Recent decades have seen a surge in interest for evidence-based policy-making. This is a welcome trend but it sets high demands for the corresponding evidence. Typical policy questions—how much will the variable of interest increase or decrease if we change a policy parameter—require estimation of causal effects that are, unfortunately, hard to identify based on commonly available data. The reasons are related to sample selection, the fact that these are typically different people and different economies that face different policy variables. For instance, workers who sign up for a training program may be more motivated or faster learners than those who do not enter the program. And if their post-program outcome differs, this may just reflect the obvious: different people behave in a different way. Unfortunately, the gold standard for measuring causal effects, randomized experiments, are sometimes too expensive or completely unfeasible.

An econometric solution to these problems is offered by Heckman (1976). The paper suggests to rephrase the model in terms of a latent variable, "participation tendency", and assumes all the disturbance terms are drawn from a common bivariate normal distribution. Although more recent literature shows that these assumptions are often unrealistic, the model remains popular in many applications due to it's simplicity and few additional demands on data. Below, we describe the model, and thereafter illustrate it's usage in `sampleSelection` package.

2 Treatment Effects with Spherical Disturbances

2.1 The Model

Assume the individual participation and outcome process is described by two latent variables: "participation tendency" y^{s*} (s stands for "selection" and asterisk * means the variable is not directly observed) and "outcome" y^o:

$$\begin{aligned} y_i^{s*} &= \alpha_0 + \boldsymbol{\alpha}_1' \boldsymbol{x}_i^s + u_i \\ y_i^o &= \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}_i^o + \beta_2 y_i^s + v_i \end{aligned} \tag{1}$$

where u and v are disturbance terms, derived from a bivariate normal distribution:

$$\begin{pmatrix} u \\ v \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma \\ \rho\sigma & \sigma^2 \end{pmatrix} \right). \tag{2}$$

$\boldsymbol{x}^s$ may include exclusion restrictions, variables not in $\boldsymbol{x}^o$, but it is not necessary as the model is also identified based on the functional form assumptions only.[1] If participation is decided based on outcome (for instance, when individuals select the training only if they expect it to pay off), one may also want that $\boldsymbol{x}^s$ to include all the components of $\boldsymbol{x}^o$. However, the general model does not require it.

Instead of the latent participation tendency we observe the actual y^s participation that occurs if $y^{s*} > 0$: $y^s = \mathbb{1}(y^{s*} > 0)$, and outcome y^o. The parameter of interest is β_2 that measures how much will y^o rise or fall if someone chooses participation instead of non-participation. Note that this specification assumes no individual heterogeneity: β_2 is constant across individuals.

Individuals participate if $y^s = \mathbb{1}(y^{s*} > 0) = 1$ i.e. $u > -\alpha_0 - \boldsymbol{\alpha}_1' \boldsymbol{x}^s$. Denote $z \equiv \alpha_0 + \boldsymbol{\alpha}_1' \boldsymbol{x}^s$ for notational simplicity, hence the participation condition can be written as $y^s = \mathbb{1}(u > -z)$. For participants

$$\mathbb{E}[y^o | \boldsymbol{x}^o, y^s = 1] = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}^o + \beta_2 + \mathbb{E}[v | u > -z] \tag{3}$$

and for non-participants

$$\mathbb{E}[y^o | \boldsymbol{x}, y^s = 0] = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}^o + \mathbb{E}[v | u < -z]. \tag{4}$$

We can identify β_2 in the usual way as $\mathbb{E}[y_i^o | \boldsymbol{x}_i, y_i^s = 1] - \mathbb{E}[y_i^o | \boldsymbol{x}_i, y_i^s = 0]$. However, as the conditional expectations in (3) and (4) are not 0, OLS estimation will give biased results.

In econometric classification, it is a switching regression (tobit-5) model where:

- Everyone has an observable outcome y^o.
- The selection indicator y^s enters the outcome equation.
- The variables $\boldsymbol{x}^o$ and parameters $\boldsymbol{\beta}_1$ are equal for both outcome types.

Note that this model cannot be estimated by the ordinary tobit-5 selection equation: intercept and β_2 are not identified unless we impose certain cross-equation restrictions. Neither can you estimate the model by tobit-2 as here both selections are observed.

2.2 Two-Step Solution

This model can be estimated by a version of Heckman (1976) two-step estimator.

First, the selection process parameters $\boldsymbol{\alpha}$ can be consistently estimated by standard probit model, and hence we can compute estimated values $\hat{z}_i$, the estimates for the true z_i.

Next, from normal density properties we know that

$$\mathbb{E}[v | u > -z] = \rho\sigma\lambda(z) \qquad \text{and} \qquad \mathbb{E}[v | u < -z] = -\rho\sigma\lambda(-z), \tag{5}$$

[1] Although formally identified, the estimates are much less precise if we do not include a strong exclusion restriction.

and

$$\sigma_0^2 \equiv \mathrm{Var}\left[v|u > -z\right] = \sigma^2 - \rho^2\sigma^2 z\lambda(z) - \rho^2\sigma^2\lambda^2(z) \tag{6}$$

$$\sigma_1^2 \equiv \mathrm{Var}\left[v|u < -z\right] = \sigma^2 + \rho^2\sigma^2 z\lambda(-z) - \rho^2\sigma^2\lambda^2(-z), \tag{7}$$

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ (commonly referred to as inverse Mill's ratio), and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions. As we have estimates for z, we can also calculate the corresponding estimates $\hat{\lambda} = \phi(\hat{z})/\Phi(\hat{z})$. Hence we can re-write the outcome equation as

$$y_i^o = \beta_0 + \boldsymbol{\beta}_1'\boldsymbol{x}_i^o + \beta_2 y_i^s + \beta_3\hat{\lambda}_i + \eta_i \tag{8}$$

where

$$\hat{\lambda}_i = \begin{cases} \lambda(z_i) & \text{if} \quad y^s = 1 \\ -\lambda(-z_i) & \text{if} \quad y^s = 0. \end{cases} \tag{9}$$

From (8) and (5) we can see that $\beta_3 = \rho\sigma$. η is a disturbance term that by construction is independent of $\hat{\lambda}$ and has variance σ_0^2 or σ_1^2, depending on the participation status. We can estimate ρ and σ from (8) in two ways. First, for participants, from (6) we have

$$\hat{\sigma}^2 = \sigma_1^2 + \rho^2\sigma^2\bar{z}\bar{\lambda}(z) + \rho^2\sigma^2\bar{\lambda}^2(z) = \sigma_1^2 + \hat{\beta}_3^2\bar{z}\bar{\lambda}(z) + \hat{\beta}_3^2\bar{\lambda}^2(z) \tag{10}$$

and second, for non-participants we get from (7)

$$\hat{\sigma}^2 = \sigma_0^2 - \rho^2\sigma^2\bar{z}\bar{\lambda}(-z) + \rho^2\sigma^2\bar{\lambda}^2(-z) = \sigma_0^2 - \hat{\beta}_3^2\bar{z}\bar{\lambda}(-z) + \hat{\beta}_3^2\bar{\lambda}^2(-z) \tag{11}$$

where upper bar denotes the corresponding sample means. σ_0^2 and σ_1^2 can be estimated from the residuals for non-participants and participants. In either case the estimator for ρ is

$$\hat{\rho} = \frac{\hat{\beta}_3}{\hat{\sigma}}. \tag{12}$$

2.3 Maximum Likelihood Estimation

It is straightforward to use Maximum Likelihood for this model. Denote the disturbance vectors by $\boldsymbol{u} = (u_1, u_2, \ldots, u_N)$ and $\boldsymbol{v} = (v_1, v_2, \ldots, v_N)$. Based on (1), the likelihood of modeled disturbances can be written as

$$\begin{aligned} \Pr(\boldsymbol{u},\boldsymbol{v}) = & \prod_{i\in\text{non-participants}} \Pr(v_i|u_i < -z_i)\Pr(u_i < -z_i)\times \\ & \times \prod_{i\in\text{participants}} \Pr(v_i|u_i > -z_i)\Pr(u_i > -z_i) \end{aligned} \tag{13}$$

Using well-known normal density properties, we get from (2):

$$\Pr(v_i|u_i < -z_i) = \frac{\frac{1}{\sigma}\phi\left(\frac{v_i}{\sigma}\right)}{\Phi(-z_i)}\Phi\left(\frac{-z_i - \frac{\rho}{\sigma}v_i}{\sqrt{1-\rho^2}}\right) \tag{14}$$

$$\Pr(u_i < -z_i) = \Phi(-z_i) \tag{15}$$

$$\Pr(v_i|u_i > -z_i) = \frac{\frac{1}{\sigma}\phi\left(\frac{v_i}{\sigma}\right)}{\Phi(z_i)}\Phi\left(-\frac{-z_i - \frac{\rho}{\sigma}v_i}{\sqrt{1-\rho^2}}\right) \tag{16}$$

$$\Pr(u_i > -z_i) = \Phi(z_i) \tag{17}$$

The disturbance terms v_i can be written based on observables as $v_i = y_i^o - \beta_0 - \boldsymbol{\beta}_1' \boldsymbol{x}_i^o - \beta_2 y_i^s$. Accordingly, we can write the model log-likelihood in the model parameters $(\alpha_0, \boldsymbol{\alpha}_1, \beta_0, \boldsymbol{\beta}_1, \beta_2, \sigma, \rho)$ and observed data $(\boldsymbol{x}, \boldsymbol{y})$ as

$$\begin{aligned}\ell = &-\frac{N}{2}\log 2\pi - N\log\sigma - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{y_i^o - \beta_0 - \boldsymbol{\beta}_1'\boldsymbol{x}_i^o - \beta_2 y_i^s}{\sigma}\right)^2 + \\ &+ \sum_{i\in\text{non-participants}} \log\Phi\left(\frac{-\alpha_0 - \boldsymbol{\alpha}_1'\boldsymbol{x}_i^s - \frac{\rho}{\sigma}\left(y_i^o - \beta_0 - \boldsymbol{\beta}_1'\boldsymbol{x}_i^o - \beta_2 y_i^s\right)}{\sqrt{1-\rho^2}}\right) + \\ &+ \sum_{i\in\text{participants}} \log\Phi\left(-\frac{-\alpha_0 - \boldsymbol{\alpha}_1'\boldsymbol{x}_i^s - \frac{\rho}{\sigma}\left(y_i^o - \beta_0 - \boldsymbol{\beta}_1'\boldsymbol{x}_i^o - \beta_2 y_i^s\right)}{\sqrt{1-\rho^2}}\right). \quad (18)\end{aligned}$$

The model is very similar in structure to the standard tobit-5 models (Amemiya, 1985; Toomet and Henningsen, 2008). Essentially it is a tobit-5 model where explanatory variables and coefficients are identical for both choices—participation and non-participation.

3 `treatReg`

3.1 Synthetic Data

Technically, `treatReg` is an amended version of tobit-5 models in the `selection` command in the package `sampleSelection2` (Toomet and Henningsen, 2008). It supports both 2-step and maximum likelihood estimation. In the latter case, 2-step method is used for calculating the initial values of parameters (unless these are supplied by the user). The only difference between `treatReg` and `selection` is the default model type: the former forces to estimate the treatment effect model, the latter detects the model type based on the arguments. If the outcome equation includes the selection outcome as an explanatory variable, it assumes the user want to treatment effect model.

First we provide an example usage using random data. We create highly correlated error terms ($\rho = 0.8$), and set all the coefficients (except the intercepts) equal to unity:

```
R> N <- 2000
R> sigma <- 1
R> rho <- 0.8
R> Sigma <- matrix(c(1, rho*sigma, rho*sigma, sigma^2), 2, 2)
R>                              # variance-covariance matrix
R> uv <- mvtnorm::rmvnorm(N, mean=c(0,0), sigma=Sigma)
R>                              # bivariate normal RV
R> u <- uv[,1]
R> v <- uv[,2]
R> x <- rnorm(N)                # normal covariates
R> z <- rnorm(N)
R> ySX <- -1 + x + z + u        # unobserved participation tendency
R> yS <- ySX > 0                # observed participation
```

```
R> yO <- x + yS + v
R> dat <- data.frame(yO, yS, x, z)
```

The code generates two correlated random variables, u and v (using `rmvnorm`). It also creates an explanatory variable x and an exclusion restriction z. Finally, we set the observable treatment indicator y^s equal to unity for those whose $y^{s*} > 0$, and calculate the outcome y^o.

First, we run a naive OLS estimate ignoring the selectivity:

```
R> m <- lm(yO ~ x + yS, data=dat)
R> print(summary(m))

Call:
lm(formula = yO ~ x + yS, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5146 -0.6649  0.0365  0.6754  2.6004

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.25460    0.02557  -9.956   <2e-16 ***
x            0.81027    0.02289  35.394   <2e-16 ***
ySTRUE       1.92569    0.05129  37.546   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9332 on 1997 degrees of freedom
Multiple R-squared:  0.7054,        Adjusted R-squared:  0.7052
F-statistic:  2391 on 2 and 1997 DF,  p-value: < 2.2e-16
```

Our estimated treatment effect (`yS`) is close to 2, instead of the correct value 1. This is because the error terms are highly positively correlated—the participants are those who have the "best" outcomes anyway. Note that the estimates for the intercept and x are biased too.

Next we use the correct statistical model with `treatReg`. We have to specify two equations: the first one is the selection equation and the second one the outcome equation. The treatment indicator enters in the latter as an ordinary control variable:

```
R> tm <- treatReg(yS ~ x + z, yO ~ x + yS, data=dat)
R> print(summary(tm))

--------------------------------------------
Tobit treatment model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
Return code 1: gradient close to zero (gradtol)
Log-Likelihood: -3254.356
2000 observations: 1419 non-participants (selection FALSE) and 581
 participants (selection TRUE)
```

```
8 free parameters (df = 1992)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.02120    0.04376  -23.34   <2e-16 ***
x            1.01973    0.04555   22.39   <2e-16 ***
z            1.05186    0.04651   22.62   <2e-16 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01565    0.02943   0.532    0.595
x            0.99599    0.02569  38.769   <2e-16 ***
ySTRUE       0.98779    0.06571  15.033   <2e-16 ***
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  1.00761    0.01888   53.38   <2e-16 ***
rho    0.81050    0.02385   33.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

The estimates are divided into three blocks: the first block describes the selection equation, the next one the outcome, and the last block describes the error terms. Note that the selection variable is listed with the corresponding factor level (here `ySTRUE`). In this case all the estimates are close to their true values. This is not surprising as we have specified the model correctly. We also recover the error term correlation 0.8 rather precisely.

3.2 Labor Market Training Data

However, the real life is almost never that simple. The data in the example above has two advantages not commonly seen in real data: first, the model is correctly specified, and second—the treatment effect is extremely strong with $\beta_2 = \sigma$, the disturbance variance in the outcome process.

Let us analyze real treatment data from library `Ecdat`. This is a US training program data from 1970s. `educ` measures education (in years), `u74` and `u75` are unemployment indicators for 1974 and 1975, `ethn` is race ("black", "hispanic" and "other") and `re78` measures real income in 1978. The logical `treat` tells if the individual was treated. First, choose `u74` and `u75` as exclusion restrictions. This amounts to assuming that previous unemployment is unrelated to the wage a few years later, except through eventual training.

```
R> data(Treatment, package="Ecdat")
R> er <- treatReg(treat~poly(age,2) + educ + u74 + u75 + ethn,
+                 log(re78)~treat + poly(age,2) + educ + ethn,
+                 data=Treatment)
R> print(summary(er))

--------------------------------------------
Tobit treatment model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 4 iterations
Return code 1: gradient close to zero (gradtol)
```

```
Log-Likelihood: -2651.502
2344 observations: 2204 non-participants (selection FALSE) and 140
 participants (selection TRUE)

17 free parameters (df = 2327)
Probit selection equation:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.94272    0.38051  -5.106 3.57e-07 ***
poly(age, 2)1 -41.64058    7.63374  -5.455 5.42e-08 ***
poly(age, 2)2   2.65968    4.97762   0.534 0.593166
educ           -0.13661    0.03207  -4.260 2.13e-05 ***
u74TRUE         0.79452    0.22374   3.551 0.000391 ***
u75TRUE         2.31494    0.21291  10.873  < 2e-16 ***
ethnblack       1.35300    0.18734   7.222 6.89e-13 ***
ethnhispanic    1.31932    0.29465   4.478 7.91e-06 ***
Outcome equation:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.983926   0.069341 129.561  < 2e-16 ***
treatTRUE     -0.963132   0.075837 -12.700  < 2e-16 ***
poly(age, 2)1  6.512273   0.797670   8.164 5.25e-16 ***
poly(age, 2)2 -4.428831   0.773235  -5.728 1.15e-08 ***
educ           0.080227   0.005231  15.338  < 2e-16 ***
ethnblack     -0.256112   0.035865  -7.141 1.23e-12 ***
ethnhispanic  -0.007786   0.079273  -0.098    0.922
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  0.69304    0.01014  68.359  < 2e-16 ***
rho    0.17699    0.06502   2.722  0.00654 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

We see that low education and unemployment are strong predictors for training participation. We also see that blacks and hispanics are more likely to be trained that "others". Surprisingly, the trainings seems to have a strong negative impact on earnings: the estimate -0.96 means that participants earn less than 40% of what the non-participants do!

Let's now acknowledge that previous unemployment may also have direct causal effect on wage and add the variables `u74` and `u75` to the outcome equation too. Now we do not have any exclusion restriction and the identification is solely based on the functional form assumptions.

```
R> noer <- treatReg(treat~poly(age,2) + educ + u74 + u75 + ethn,
+                   log(re78)~treat + poly(age,2) + educ + u74 + u75 + ethn,
+                   data=Treatment)
R> print(summary(noer))

--------------------------------------------
Tobit treatment model (switching regression model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 3 iterations
```

```
Return code 1: gradient close to zero (gradtol)
Log-Likelihood: -2613.995
2344 observations: 2204 non-participants (selection FALSE) and 140
 participants (selection TRUE)

19 free parameters (df = 2325)
Probit selection equation:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.93285    0.38110  -5.072 4.25e-07 ***
poly(age, 2)1 -42.90457    7.99609  -5.366 8.86e-08 ***
poly(age, 2)2   0.95030    5.15903   0.184  0.85387
educ           -0.13664    0.03209  -4.258 2.14e-05 ***
u74TRUE         0.70914    0.21806   3.252  0.00116 **
u75TRUE         2.27799    0.20967  10.865  < 2e-16 ***
ethnblack       1.31536    0.18566   7.085 1.84e-12 ***
ethnhispanic    1.26579    0.29817   4.245 2.27e-05 ***
Outcome equation:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.996364   0.068110 132.086  < 2e-16 ***
treatTRUE     -0.508259   0.106089  -4.791 1.77e-06 ***
poly(age, 2)1  7.026173   0.786638   8.932  < 2e-16 ***
poly(age, 2)2 -4.701016   0.761097  -6.177 7.71e-10 ***
educ           0.080785   0.005141  15.714  < 2e-16 ***
u74TRUE       -0.580994   0.071644  -8.109 8.14e-16 ***
u75TRUE       -0.030988   0.083291  -0.372    0.710
ethnblack     -0.269380   0.035322  -7.626 3.49e-14 ***
ethnhispanic  -0.004216   0.077958  -0.054    0.957
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  0.68058    0.00994  68.466   <2e-16 ***
rho   -0.02145    0.06733  -0.319     0.75
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

Now the estimated treatment effect is substantially smaller in absolute value, only -0.51, and hence participants earn about 60% of income of non-participants.

We also see that while the error terms in the first model above were slightly positively correlated, now these are essentially independent. However, as the selection equation estimates suggest, the participants are drawn from the weak end of the observable skill distribution. If this is also true for unobservables, we would expect the correlation to be negative. Seems like this data is too coarse to correctly determine the bias. The reader is encouraged to experiment with further variables in the data, such as pre-program incomes.

4 Conclusion

Treatment effect models with spherical disturbances remain popular in applied research despite the often disputed assumptions. `sampleSelection` offers an easy interface to estimate such models.

References

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

Heckman, J. J. (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**, 475–492.

Toomet, O. and Henningsen, A. (2008) Sample selection models in R: Package sampleSelection, *Journal of Statistical Software*, **27**.