Treatment Effect with Normal Disturbances: `treatReg`

Ott Toomet
Tartu University

October 22, 2017

1 The Problem

The holy grail of policy analysis is to determine causal effects. Causal parameters can be directly interpreted as "impact": how much does the variable of interest increase if we change a policy parameter? Such effects are hard to estimate based on commonly available data. The reason is self-selection, the fact that these are typically different people who face different policy variables. If their outcome-of-interest differs, this just can reflect the obvious: different people behave in a different way. Unfortunately, the gold standard for determining causality, randomized experiment, is too often not feasible either. A solution is offered by Heckman (1976). That paper proposes 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.

Assume two underlying latent variables: "participation tendency" y^{s*} 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 $y^s = \mathbb{1}(y^{s*} > 0)$ is the observable participation indicator and u and v are normally distributed disturbance terms:

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

$\boldsymbol{x}^s$ may include variables, not in $\boldsymbol{x}^o$ (exclusion restrictions) but it is not necessary. The parameter of interest is β_2. We observe the actual participation y^s and the outcome $y^o = y^{o*}$.

Individuals participate if $y^{s*} > 0$ i.e. $u > -\alpha_0 - \boldsymbol{\alpha}_1' \boldsymbol{x}^s$ and hence for participants

$$\mathbb{E}[y^o | \boldsymbol{x}^o, y_i^s = 1] = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}^o + \beta_2 + \mathbb{E}[v | u > -\alpha_0 - \boldsymbol{\alpha}_1' \boldsymbol{x}^s] \tag{3}$$

and for non-participants

$$\mathbb{E}[y_i^o | \boldsymbol{x}_i, y_i^s = 0] = \beta_0 + \boldsymbol{\beta}_1' \boldsymbol{x}^o + \mathbb{E}[v | u < -\alpha_0 - \boldsymbol{\alpha}_1' \boldsymbol{x}^s]. \tag{4}$$

We can identify β_2 in the usual way as $\mathbb{E}[y_i^o | \boldsymbol{x}_i, y_i^s = 1] - \mathbb{E}[y_i^o | \boldsymbol{x}_i, y_i^s = 0]$.

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

- Everyone has an observable outcome y^o.
- There is an selection indicator y^s in the outcome equation.
- The variables $\boldsymbol{x}$ and parameters α_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 Two-Step Solution

This model can be consistently estimated by a version of Heckman (1976) two-step estimator. First, one can consistently estimate the selection process by probit.

Next, denote $z = \alpha_0 + \boldsymbol{\alpha}_1' \boldsymbol{x}^s$. From normal density properties we know that

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

and

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

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

where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$, and ϕ and Φ are standard normal pdf and cdf correspondingly. 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

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

η is a disturbance term that by construction is independent of $\hat{\lambda}$ and has variance of χ_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 mean. σ_0^2 and σ_1^2 can simply be estimated from the residuals for non-participants and participants. In both case the estimator for ρ is

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

3 Maximum Likelihood Estimation

Denote by $\boldsymbol{u} = (u_1, u_2, \ldots, u_N)$ and $\boldsymbol{v} = (v_1, v_2, \ldots, v_N)$. Based on (1) we can write

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

Normal density properties tell that

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

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

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

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

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions. 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 as

$$\begin{aligned}\ell = -\frac{N}{2}\log 2\pi - N\log\sigma - \frac{N}{2}\left(\frac{v_i}{\sigma}\right)^2 + \\ + \sum_{i\in\text{non-participants}} \log\Phi\left(\frac{-z_i - \frac{\rho}{\sigma}v_i}{\sqrt{1-\rho^2}}\right) + \\ + \sum_{i\in\text{non-participants}} \log\Phi\left(-\frac{-z_i - \frac{\rho}{\sigma}v_i}{\sqrt{1-\rho^2}}\right).\end{aligned} \tag{18}$$

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

4 `treatReg`

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 nitial values of parameters (unless these are supplied by the user).

We first provide a random data example. We use highly correlated error terms ($\rho = 0.8$), all the coefficients are 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> uv <- rmvnorm(N, mean=c(0,0), sigma=Sigma)
R> u <- uv[,1]
R> v <- uv[,2]
R> x <- rnorm(N)
R> z <- rnorm(N)
R> ySX <- -1 + x + z + u
R> yS <- ySX > 0
R> yO <- x + yS + v
R> dat <- data.frame(yO, yS, x, z, ySX, u, v)
```

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 x^s equal to unity for those whose $x^{s*} > 0$, and calculate the outcome y^o.

First, we run a naive OLS estimate completely 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 also the estimates for the intercept and x are wrong.

Now we estimate the same problem using the correct statistical model with `treatReg`. We have to specify two equations: the first one is the selection equation, the second one the actual outcome. The treatment indicator enters here 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
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. In this case all the estimates are close to their true values. This is not surprising we have specified the model correctly. We also rather precisely recover the error term correlation 0.8, note that it also possesses extremely high t-value of 33.

However, the real life is almost always harder. The example above involves two advantages not commonly seen in real data: first, the model is correctly specified, and second—the treatment effect is extremely strong with $\beta_2 = \sigma$.

Let's analyze a real dataset (treatment data from library `Ecdat`). This includes 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. 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
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.

```
R> ## The treatment effect estimate 'treatTRUE' is -0.96, i.e. the
R> ## treatment substantially lower the earnings
R> ## Now estimate it withouth the exclusion restriction
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
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 the error terms in the first case are slightly positively correlated while in the latter case they are essentially uncorrelated. However, as the selection equation estimates suggest, the participants are drawn from the weakest 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.

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**.