POLS6382 Quantitative Research Methods III: Maximum Likelihood Estimation

Ling Zhu
Department of Political Science
University of Houston
October 18, 2021

Lab 7: Models for Censored and Truncated Data

Objectives

- Compare Tobit and OLS with censored data.
- Learn how to estimate a Heckman selection model.

1 Compare Tobit and OLS with Censored Data

In this section, we will use a simulated dataset to compare Tobit and OLS model. First, we simulate a dataset that contains a dependent variable y, which is left-censored. All the censored cases are assigned with value "0". The explanatory variable is x, which is generated by randomly drawing numbers from a normal distribution. We also set the latent variable y.star as a linear function of x, with a slope coefficient of 5.

———————————- R Code———————————-

```
N = 10
f = rep(c("s1","s2","s3","s4","s5","s6","s7","s8"),N)
fcoeff = rep(c(-1,-2,-3,-4,-3,-5,-10,-5),N)
set.seed(100)
x = rnorm(8*N)+1
beta = 5
epsilon = rnorm(8*N,sd = sqrt(1/5))
y.star = x*beta+fcoeff+epsilon ## latent response
y = y.star
y[y<0] <- 0 ## censored response
simdata<-data.frame(cbind(x,y))
```

———————————————————————————————-

Next, we fit an OLS model with this simulated dataset. We see that the OLS model produces a slope coefficient of 2.892, which is substantially smaller than the true parameter value that we set (β=5). The biased coefficient is caused by ignoring the fact that values of y are censored at 0.

———————————- R Code/Output———————————-

```
fitols<-lm(y~x)
summary(fitols)
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.1985 -1.3461  0.0779  1.1305  5.9851

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  -0.1398     0.3155  -0.443    0.659
x             2.8926     0.2178  13.280   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.084 on 78 degrees of freedom
Multiple R-squared:  0.6934,Adjusted R-squared:  0.6894
F-statistic: 176.4 on 1 and 78 DF,  p-value: < 2.2e-16
```

The proper model specification is Tobit. We use the `vglm` function to estimate a Tobit model.[1]. The first coefficient labeled as "(intercept):1" is the intercept term. The second coefficient labeled as "(intercept):2" an ancillary statistic. If we exponentiate this value, we get a number that is analogous to the square root of the residual variance in OLS regression. $exp(-1.104) \approx 0.3315$, which is substantially smaller than the residual variance of the OLS model. We also see that the Tobit model produces a mean slope of 4.758 for variable x. This is quite close to the true parameter value (5.)

---- R Code/Output----

```
fittobit<-vglm(formula=y~x, family=tobit(Lower=0))
summary(fittobit)
Call:
vglm(formula = y ~ x, family = tobit(Lower = 0))

Pearson residuals:
            Min      1Q  Median        3Q   Max
mu       -2.729 -0.4390 -0.0145  0.680031 1.771
loge(sd) -1.104 -0.5756 -0.1882 -0.004532 4.610

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -3.6391     0.5886  -6.182 6.31e-10 ***
(Intercept):2   0.7693     0.1064   7.228 4.89e-13 ***
x               4.7582     0.3438  13.840  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Number of linear predictors:  2
Names of linear predictors: mu, loge(sd)
Dispersion Parameter for tobit family:   1
Log-likelihood: -110.6193 on 157 degrees of freedom
Number of iterations: 5
```

We can also compare the two models by plotting the observations and the two regression lines. Figure 1 shows the Tobit model (orange line) and the OLS model (red line) produces very different slope coefficients. Ignoring that y is censored, the OLS model underestimates the slope coefficient of x.

---- R Code----

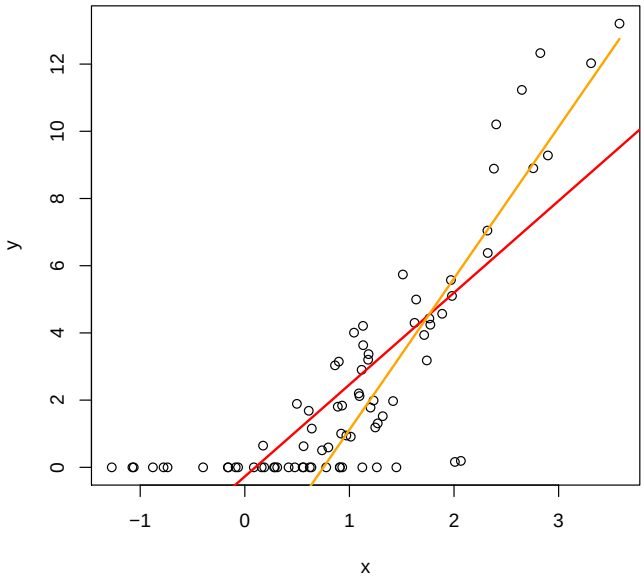
```
pdf(file="olstobit.pdf", width=6, height=6)
```

[1]Various `R` functions can be used to estimate a Tobit model, e.g. `censReg()` from package `censReg` and `tobit()` from package `AER`.

```
plot(x,y)
abline(lm(y~x),col="red",lwd=2,lty=1)
curve(-3.3862557 + 4.5046179 *x, lty="orange", lwd="2", add=TRUE)
dev.off()
```

- R Output

Figure 1: Comparing OLS and Tobit with Left-Censored Data



2 Tobit Model for Left- and Right-Censored Data

We use a dataset named `tobit.csv`. This dataset considers the situation in which we have a measure of academic aptitude (scaled 200-800) which we want to model using reading and math test scores, as well as the type of program the student is enrolled in (academic, general, or vocational). The problem here is that students who answered all questions on the academic aptitude test correctly receive a score of 800, even though it is likely that these students are not "truly" equal in aptitude. The same is true of students who answer all of the questions incorrectly. All such students would have a score of 200, although they may not all be of equal aptitude.

The dataset contains 200 observations. The academic aptitude variable is `apt`, the reading and math test scores are `read` and `math` respectively. The variable `prog` is the type of program the student is in, it is a categorical (nominal) variable that takes on three values, academic (`prog` = 1), general (`prog` = 2), and vocational (`prog` = 3). The variable `id` indexes different students.

Let us start from looking at the dependent variable `apt` descriptively. Note that in this dataset, the lowest value of `apt` is 352. That is, no students received a score of 200 (the lowest score possible), meaning that even though censoring from below was possible, it does not occur in the dataset.

- R Code/Output

```
summary(mydata$apt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  352.0   575.5   633.0   640.0   705.2   800.0
```

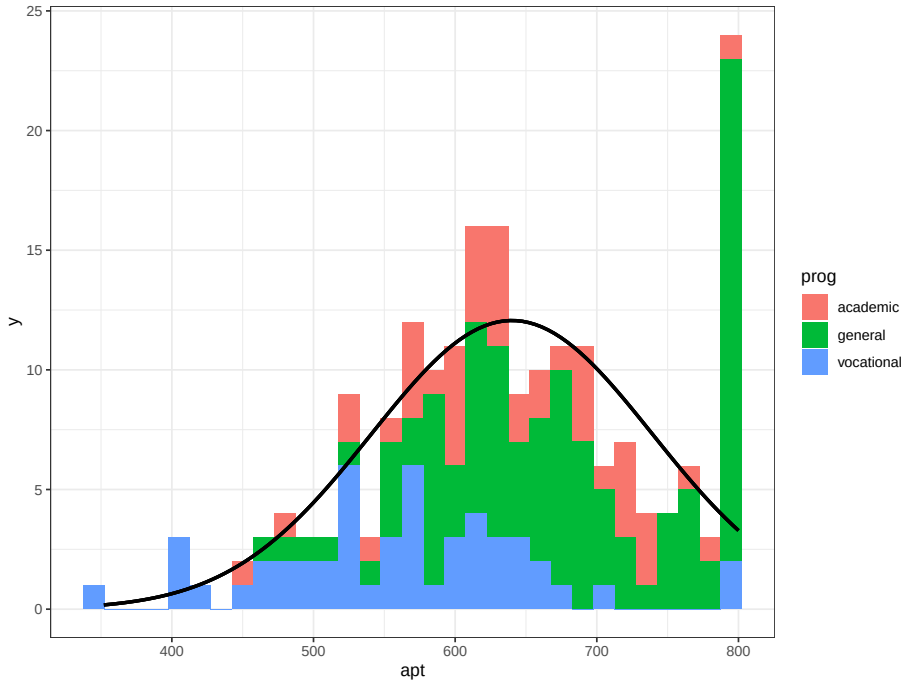
Figure 2 is a histogram plot, showing the frequency of different scores. Figure 2 clearly shows that the censoring in the values of `apt`. There are far more cases with scores of 750 to 800 than one would expect looking at the rest of the distribution.

- R Code

```
f <- function(x, var, bw = 15) {
  dnorm(x, mean = mean(var), sd(var)) * length(var)  * bw
}
pdf(file="apt.pdf",height=6, width=6)
p<-ggplot(mydata, aes(x = apt, fill=prog))
p + stat_bin(binwidth=15) +
  stat_function(fun = f, size = 1,
  args = list(var = mydata$apt))
  dev.off()
```

- R Output

Figure 2: Histogram of Variable `apt` (Academic Aptitude)



We can also describe variables by using a correlation matrix figure. Figure 3 shows both `read` and `math` are positively correlated with `apt`.

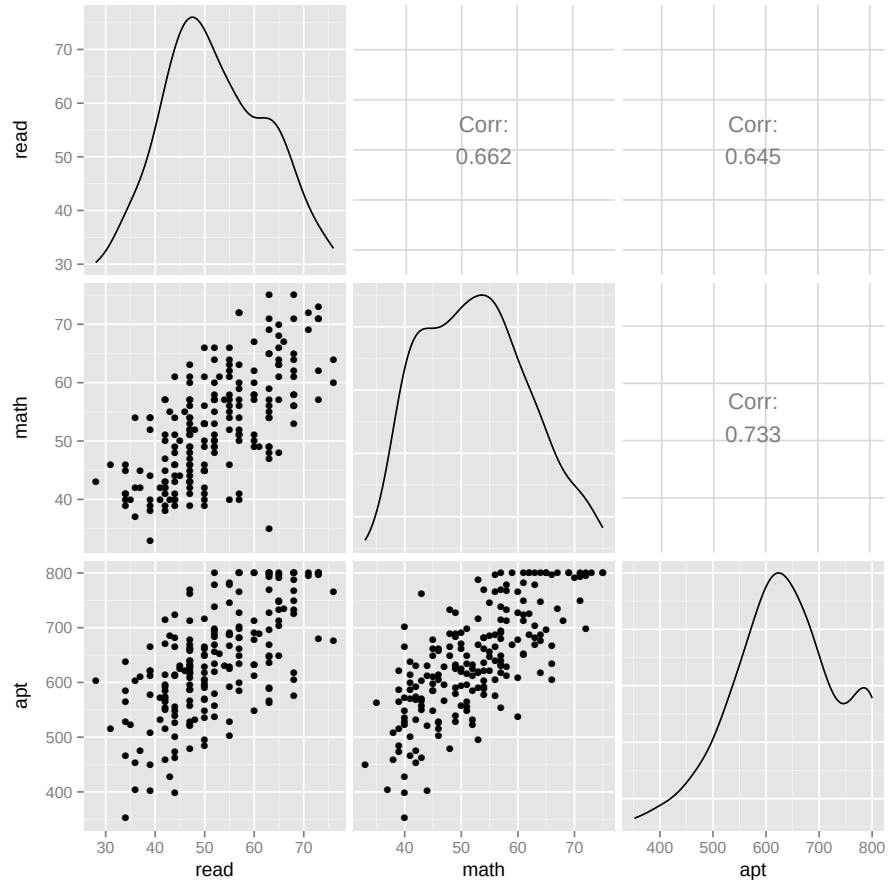
- R Code

```
cor(mydata[, c("read", "math", "apt")])
```

```
# Plot matrix
pdf(file="correlation.pdf",height=8, width=8)
ggpairs(mydata[, c("read", "math", "apt")])
dev.off()
```

- R Output -

Figure 3: Correlation Matrix



Below, we run a tobit model, using the `vglm` function from the VGAM package.

- R Code/Output -

```
tobitmodel<-vglm(formula=apt~read+math+as.factor(prog),
                 family=tobit(Upper=800))
summary(tobitmodel)

Call:
vglm(formula = apt ~ read + math + as.factor(prog), family = tobit(Upper = 800))

Pearson residuals:
            Min      1Q   Median     3Q   Max
```

```
mu       -2.656 -0.7764 -0.04379 0.7514 4.202
loge(sd) -1.406 -0.6212 -0.32468 0.2547 5.188

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept):1             209.56587   32.74635   6.400 1.56e-10 ***
(Intercept):2               4.18474    0.05278  79.294  < 2e-16 ***
read                        2.69794    0.62064   4.347 1.38e-05 ***
math                        5.91449    0.70628   8.374  < 2e-16 ***
as.factor(prog)general    -12.71475   12.38546  -1.027  0.30461
as.factor(prog)vocational -46.14390   13.74876  -3.356  0.00079 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Number of linear predictors:  2
Names of linear predictors: mu, loge(sd)
Dispersion Parameter for tobit family:   1
Log-likelihood: -1041.063 on 394 degrees of freedom
Number of iterations: 5
```

Based on this model, we can interpret the results substantively as the following:

- For a one unit increase in `read`, there is a 2.6981 point increase in the predicted value of `apt`.
- A one unit increase in `math` is associated with a 5.9146 unit increase in the predicted value of `apt`.
- The terms for `prog` have a slightly different interpretation. The predicted value of `apt` is 46.1419 points lower for students in a vocational program than for students in an academic program.
- We do not observe statistically different predicted `apt` scores between students in a general program and those in an academic program.

We can test the significant of program type overall by fitting a model without variable " program" in it and using the likelihood ratio test to compare two models. The LRT with two degrees of freedom is associated with a p-value of 0.0032, indicating that the overall effect of `prog` is statistically significant.

———————————- R Code/Output————————————————————————-

```
tobitmodel2<-vglm(apt ~ read + math, tobit(Upper = 800),
                  data = mydata)
(p <- pchisq(2 * (logLik(tobitmodel) - logLik(tobitmodel2)),
             df = 2, lower.tail = FALSE))
[1] 0.003155176
```

———-

3 Heckman Sample Selection Models: Simulated Data

In this section, we use a simulated dataset to show how we can estimate a correctly specified Heckman selection model with exclusion restriction. We simulate the data by following steps.

- Using `mvtnorm`, we create bivariate normal disturbances with correlation -0.7. This is the correlation parameter ρ between our selection and outcome equation.
- We generate a uniformly distributed explanatory variable for the selection equation, `xs`, the selection outcome `ys` by Probit data generating process,
- The explanatory variable for the outcome equation `xo` is also drawn from a uniform distribution.
- All our true intercepts are equal to 0 and our true slopes are equal to 1, both in this and the following examples.
- The latent outcome variable is `yoX`, and the observable outcome is `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.

———————————- R Code————————————————————————————————-

```
set.seed(0)
eps <- rmvnorm(500, c(0, 0),
     matrix(c(1, -0.7, -0.7, 1), 2, 2))
# selection
xs <- runif(500)
ys <- xs + eps[, 1] > 0
# outcome
xo <- runif(500)
yoX <- xo + eps[, 2]
yo <- yoX * (ys > 0)
```

Next, we run a Heckman selection model using function `selection` from `sampleSelection` package. The first model component is the selection equation, and the second component is the outcome equation. We see that the estimates are reasonably precise.

———————————- R Code————————————————————————————————-

```
summary(selection(ys ~ xs, yo ~ xs))
--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 5 iterations
Return code 1: gradient close to zero
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   0.0393 *
xs            1.3377     0.2014   6.642 3.09e-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 8.13e-06 ***
Error terms:
      Estimate Std. error t value  Pr(> t)
sigma   0.9190     0.0574  16.009  < 2e-16 ***
rho    -0.5392     0.1521  -3.544 0.000394 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

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

——————————- R Code ——————————————————————————-

```
yoX <- xs + eps[, 2]
yo <- yoX * (ys > 0)
summary(selection(ys ~ xs, yo ~ xs))
--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 15 iterations
Return code 1: gradient close to zero
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   0.0749 .
xs            1.2907     0.2085   6.191 5.96e-10 ***
Outcome equation:
            Estimate Std. error t value Pr(> t)
(Intercept)  -0.5500     0.5645  -0.974 0.32997
xs            1.3987     0.4483   3.120 0.00181 **
Error terms:
      Estimate Std. error t value Pr(> t)
sigma  0.85091    0.05352  15.899  <2e-16 ***
rho   -0.13223    0.72698  -0.182   0.856
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```

4 Heckman Selection Model with Observational Data

The data set used in this example is included in `sampleSelection`, named `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 estimate the model by the Heckman two-step method.

——————————- R Code/Output ——————————————————————————-

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

# Heckman Two-Step Method
```

```
selectmod1 <- selection(lfp ~ age + I(age^2) + faminc + kids + educ,
  +   wage ~ exper + I(exper^2) + educ + city, data = Mroz87,
    method = "2step")
--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
753 observations (325 censored and 428 observed)
14 free parameters (df = 740)
Probit selection equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.157e+00  1.402e+00  -2.965 0.003127 **
age          1.854e-01  6.597e-02   2.810 0.005078 **
I(age^2)    -2.426e-03  7.735e-04  -3.136 0.001780 **
faminc       4.580e-06  4.206e-06   1.089 0.276544
kidsTRUE    -4.490e-01  1.309e-01  -3.430 0.000638 ***
educ         9.818e-02  2.298e-02   4.272 2.19e-05 ***
Outcome equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9712003  2.0593505  -0.472    0.637
exper        0.0210610  0.0624646   0.337    0.736
I(exper^2)   0.0001371  0.0018782   0.073    0.942
educ         0.4170174  0.1002497   4.160 3.56e-05 ***
city         0.4438379  0.3158984   1.405    0.160
Multiple R-Squared:0.1264,Adjusted R-Squared:0.116
Error terms:
             Estimate Std. Error t value Pr(>|t|)
invMillsRatio   -1.098      1.266  -0.867    0.386
sigma            3.200         NA      NA       NA
rho             -0.343         NA      NA       NA
--------------------------------------------
```

In this exercise, we are interested in examining the determinants of female workers' wage. To do so, we have to consider the selection process that determines female labor participation. The above model shows that education affects both labor participation and wage. Notice that the above model also produces "NAs" for sigma and ρ, which means that we have a non positive and definite variance covariance matrix. This could be because that we add quadratic polynomial terms in both equations. When this issue occurs, we can consider the ML estimation

——————————- R Code/Output———————————————————————————-

```
selectmod2 <- selection(lfp ~ age + I(age^2) + faminc + kids + educ,
          +wage ~ exper + I(exper^2) + educ + city, data = Mroz87,
          maxMethod = "BHHH", iterlim = 500)
summary(selectmod2)
--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
BHHH maximisation, 144 iterations
Return code 2: successive function values within tolerance limit
Log-Likelihood: -1581.258
753 observations (325 censored and 428 observed)
13 free parameters (df = 740)
Probit selection equation:
```

```
              Estimate Std. error t value  Pr(> t)
(Intercept) -4.120e+00  1.410e+00  -2.921 0.003491 **
age          1.840e-01  6.584e-02   2.795 0.005193 **
I(age^2)    -2.409e-03  7.735e-04  -3.114 0.001845 **
faminc       5.680e-06  3.889e-06   1.460 0.144217
kidsTRUE    -4.506e-01  1.367e-01  -3.297 0.000977 ***
educ         9.528e-02  2.400e-02   3.970 7.18e-05 ***
Outcome equation:
             Estimate Std. error t value  Pr(> t)
(Intercept) -1.962922   1.680266  -1.168    0.243
exper        0.027874   0.075611   0.369    0.712
I(exper^2)  -0.000104   0.002341  -0.044    0.965
educ         0.456997   0.096268   4.747 2.06e-06 ***
city         0.446514   0.426922   1.046    0.296
Error terms:
      Estimate Std. error t value Pr(> t)
sigma  3.10832    0.08364   37.16  <2e-16 ***
rho   -0.13197    0.22377   -0.59   0.555
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
--------------------------------------------
```