Econometrics II TA Session #5

Hiroki Kato

1 Empirical Application of Truncated Regression: Labor Participation of Married Women (1)

1.1 Background and Data

To develop women's social advancement, we should create environment to keep a good balance between work and childcare after marriage. In this application, using the dataset of married women, we explore how much childcare prevents married women to participate in labor market.

Our dataset originally comes from Stata sample data. [1] This dataset contains the following variables:

- `whrs`: Hours of work. This outcome variable is truncated from below at zero.
- `kl6`: the number of preschool children
- `k618`: The number of school-aged children
- `wa`: age
- `we`: The number of years of education

```
dt <- read.csv(file = "./data/labor.csv", header = TRUE,  sep = ",")
summary(dt)
```

```
##       whrs           kl6              k618              wa
##  Min.   :  12   Min.   :0.0000   Min.   :0.000   Min.   :30.00
##  1st Qu.: 645   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:35.00
##  Median :1406   Median :0.0000   Median :1.000   Median :43.50
##  Mean   :1333   Mean   :0.1733   Mean   :1.313   Mean   :42.79
##  3rd Qu.:1903   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:48.75
##  Max.   :4950   Max.   :2.0000   Max.   :8.000   Max.   :60.00
##        we
##  Min.   : 6.00
##  1st Qu.:12.00
##  Median :12.00
```

[1] http://www.stata-press.com/data/r13/laborsub.dt. Because this is dta file, we need to import it, using the `read.dta` function in the library `foreign`. I intentionally remove married women who could not participate in the labor market.

```
## Mean   :12.64
## 3rd Qu.:13.75
## Max.   :17.00
```

1.2 Model

Since we cannot observe those who could not partiapte in the labor market (`whrs = 0`), we use the truncated regression model. Thus, the selection rule is as follows:

$$\begin{cases} y_i = \mathbf{x}_i\beta + u_i & \text{if } s_i = 1 \\ s_i = 1 & \text{if } a_1 < y_i < a_2 \end{cases}.$$

where $u_i \sim N(0, \sigma^2)$. By the distributional assumption, we have $y_i|\mathbf{x}_i \sim N(\mathbf{x}_i\beta, \sigma^2)$. In this application, we set $a_1 = 0$ and $a_2 = +\infty$.

Since we are interested in estimating β, we must condition on $s_i = 1$. The probability density function of y_i conditional on $(x_i, s_i = 1)$ is

$$p_\theta(y_i|\mathbf{x}_i, s_i = 1) = \frac{f(y_i|\mathbf{x}_i)}{\mathbb{P}(s_i = 1|\mathbf{x}_i)}.$$

where $\theta = (\beta, \sigma^2)'$. By the distributional assumption, the conditional distribution of y_i is given by

$$f(y_i|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right)^2\right) = \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right),$$

where $\phi(\cdot)$ is the standard normal density function. Moreover, the probability of observation $(s_i = 1)$ is given by

$$\begin{aligned} \mathbb{P}(s_i = 1|\mathbf{x}_i) &= \mathbb{P}(\mathbf{x}_i\beta + u_i > 0|\mathbf{x}_i) \\ &= \mathbb{P}(u_i/\sigma > -\mathbf{x}_i\beta/\sigma|\mathbf{x}_i) \\ &= 1 - \Phi\left(\frac{y_i - \mathbf{x}_i\beta}{\sigma}\right), \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative density function.

Thus, the log-likelihood function is

$$M_n(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma}\frac{\phi(\frac{y_i - x_i\beta}{\sigma})}{1 - \Phi(\frac{-x_i\beta}{\sigma})}\right).$$

We provide two ways to estimate truncated regression, using `R`. First way is to define the log-likelihood function directly and minimize its function by `nlm` function. Recall that `nlm` function provides the Newton method to minimize the function. We need to give intial values in argument of this function. To set initial values, we assume that coefficients of explanatory variables are zero. Then, we obtain $y_i|\mathbf{x}_i \sim N(\beta_1, \sigma^2)$. Thus, the initial value of σ, `b[1]` is the standard deviation of `whrs`, and the initial value of β_1, `b[2]` is the mean of `whrs`. Note that these initial values are not unbised estimator.

```
whrs <- dt$whrs
kl6 <- dt$kl6; k618 <- dt$k618
wa <- dt$wa; we <- dt$we

LnLik <- function(b) {
  sigma <- b[1]
  xb <- b[2] + b[3]*kl6 + b[4]*k618 + b[5]*wa + b[6]*we
  condp <- dnorm((whrs - xb)/sigma)/(1 - pnorm(-xb/sigma))
  LL_i <- log(condp/sigma)
  LL <- -sum(LL_i)
  return(LL)
}

init <- c(sd(whrs), mean(whrs), 0, 0, 0, 0)
est.LnLik <- nlm(LnLik, init, hessian = TRUE)
```

Second way is to use the function `truncreg` in the library `truncreg`. We must specify the trucated point, using arguments `point` and `direction`. If `direction = "left"`, the outcome variable is truncated from below at `point`, that is, `point < y`. On the other hand, if `direction = "right"`, the outcome variable is truncated from above at `point`, that is, `y < point`.

```
library(truncreg)
model <- whrs ~ kl6 + k618 + wa + we
est.trunc <- truncreg(model, data = dt, point = 0, direction = "left")
se.trunc <- sqrt(diag(vcov(est.trunc)))
```

1.3 Interpretations

Table 1 shows results of truncated regression estimated by two methods. As a comparison, we also show the OLS result in column (3). All specifications show that the number of preschool and school-aged children reduces the hours of work. The size of coefficient of the number of preschool and school-aged children become stronger when we use the truncated regression. Note that the size of coeffieient of `#.Preschool Children` estimated by `truncreg` is different from the coefficient estimated by `nlm`.

```
ols <- lm(model, data = dt)
coef.LnLik <- est.LnLik$estimate
se.LnLik <- sqrt(diag(solve(est.LnLik$hessian)))
names(coef.LnLik) <- c("sigma", names(coef(ols)))
```

```r
names(se.LnLik) <- c("sigma", names(coef(ols)))

library(stargazer)
stargazer(
  ols, ols, ols,
  column.labels = c("Truncated (truncreg)", "Truncated (nlm)", "OLS"),
  coef = list(coef(est.trunc), coef.LnLik[2:6]),
  se = list(se.trunc, se.LnLik[2:6]),
  report = "vcs", keep.stat = c("n"),
  covariate.labels = c(
    "\\#.Preschool Children",
    "\\#.School-aged Children",
    "Age", "Education Years"
  ),
  add.lines = list(
    c("Estimated Sigma",
      round(coef(est.trunc)[6], 3), round(coef.LnLik[1], 3)),
    c("Log-Likelihood",
      round(est.trunc$logLik, 3), round(-est.LnLik$minimum, 3))
  ),
  omit.table.layout = "n", table.placement = "t",
  title = "Truncated Regression: Labor Market Participation of Married Women",
  label = "lfp",
  type = "latex", header = FALSE
)
```

2 Empirical Application of Tobit Regression: Labor Participation of Married Women (2)

2.1 Background and Data

We continue to investigate the previous research question. We use dataset coming from same source as the previous one. Unlike the previous dataset, we now observe married woment who do not participate in the labor market (`whrs = 0`). Additionally, we introduce the new variable:

- `lfp`: a dummy variable taking 1 if observed unit works.

The previous dataset contains observations with `lfp = 1`. In this application, we use observations with `lfp = 0` to estimate the tobit model.

```r
dt <- read.csv(file = "./data/labor2.csv", header = TRUE,  sep = ",")
summary(dt)
```

```
##       lfp            whrs              kl6              k618              wa
```

Table 1: Truncated Regression: Labor Market Participation of Married Women

	Dependent variable:		
	whrs		
	Truncated (truncreg)	Truncated (nlm)	OLS
	(1)	(2)	(3)
#.Preschool Children	−456.785	−803.032	−421.482
	(266.367)	(252.803)	(167.973)
#.School-aged Children	−153.347	−172.875	−104.457
	(81.780)	(100.590)	(54.186)
Age	−5.379	−8.821	−4.785
	(13.492)	(14.646)	(9.691)
Education Years	−0.092	16.529	9.353
	(43.702)	(46.430)	(31.238)
Constant	1,624.584	1,586.228	1,629.817
	(857.730)	(932.878)	(615.130)
Estimated Sigma	941.464	983.736	
Log-Likelihood	-1201.698	-1200.916	
Observations	150	150	150

```
##  Min.   :0.0   Min.   :    0.0   Min.   :0.000   Min.   :0.000   Min.   :30.00
##  1st Qu.:0.0   1st Qu.:    0.0   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:35.00
##  Median :1.0   Median : 406.5    Median :0.000   Median :1.000   Median :43.00
##  Mean   :0.6   Mean   : 799.8    Mean   :0.236   Mean   :1.364   Mean   :42.92
##  3rd Qu.:1.0   3rd Qu.:1599.8    3rd Qu.:0.000   3rd Qu.:2.000   3rd Qu.:49.00
##  Max.   :1.0   Max.   :4950.0    Max.   :3.000   Max.   :8.000   Max.   :60.00
##        we
##  Min.   : 5.00
##  1st Qu.:12.00
##  Median :12.00
##  Mean   :12.35
##  3rd Qu.:13.00
##  Max.   :17.00
```

2.2 Model

Our dependent variable is censored from below at zero. The censored data is caused by the corner solution problem. Married women chooses zero labor time if, without any constraint, their optimal labor time is negative. In this case, we should use the tobit model. The tobit model is

$$y_i = \begin{cases} \mathbf{x}_i \beta + u_i & \text{if } y_i > a \\ a & \text{otherwise} \end{cases},$$

where $E(u_i) = 0$ and $\text{Var}(u_i) = 0$. In this application, we set $a = 0$.

Using this model, the probability of y_i conditional on x_i is defined by

$$p_{\beta,\sigma^2}(y_i|x_i) = \mathbb{P}(y_i \leq 0)^{1[y_i=0]} f(y_i|\mathbf{x}_i)^{1-1[y_i=0]}$$

where $f(y_i|x_i)$ is the probability density function conditional on $\mathbf{x}_i$, $1[y_i = 0]$ is an indicator function returing 1 if $y_i = 0$. Now, we assume the distribution $u_i|\mathbf{x}_i \sim N(0, \sigma^2)$. Then, we can reformulate $\mathbb{P}(y_i \leq 0)$ as follows:

$$\mathbb{P}(y_i \leq 0) = \mathbb{P}(-\mathbf{x}_i \beta \leq u_i) = \Phi\left(-\frac{\mathbf{x}_i \beta}{\sigma}\right) = 1 - \Phi\left(\frac{\mathbf{x}_i \beta}{\sigma}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the stadnard normal distribution. Note that the last equatility comes from symmetric property of the standard normal distribution. Moreover, the density function f is reformulated as follows:

$$f(y_i|\mathbf{x}_i) = \frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i \beta}{\sigma}\right).$$

Assuming iid sample, we obtain the join probability function as follows:

$$p_{\beta,\sigma^2}((y_i|x_i), i = 1, \dots, n) = \prod_{i=1}^{n} \left(1 - \Phi\left(\frac{\mathbf{x}_i \beta}{\sigma}\right)\right)^{1[y_i=0]} \left(\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i \beta}{\sigma}\right)\right)^{1-1[y_i=0]}.$$

We estimate $\log p_{\beta,\sigma^2}((y_i|x_i), i = 1, \dots, n)$, using the maximum likelihood method. In `R`, there are two ways to implement the tobit regression. First way is to define the log-likelihood function directly and minimize its function by `nlm` function. We need to give intial values in argument of this function. To set initial values, we assume coefficients of explanatory variables are zero. Then, we obtain $y_i|\mathbf{x}_i \sim N(\beta_1, \sigma^2)$ where β_1 is intercept of regression equation. Thus, the initial value of σ, `b[1]` is the standard deviation of `whrs`, and the initial value of β_1, `b[2]` is the mean of `whrs`.

```r
whrs <- dt$whrs
kl6 <- dt$kl6; k618 <- dt$k618
wa <- dt$wa; we <- dt$we

LnLik <- function(b) {
  sigma <- b[1]
  xb <- b[2] + b[3]*kl6 + b[4]*k618 + b[5]*wa + b[6]*we
  Ia <- ifelse(whrs == 0, 1, 0)
  F0 <- 1 - pnorm(xb/sigma)
  fa <- dnorm((whrs - xb)/sigma)/sigma
  LL_i <- Ia * log(F0) + (1 - Ia) * log(fa)
  LL <- -sum(LL_i)
  return(LL)
}

init <- c(sd(whrs), mean(whrs), 0, 0, 0, 0)
est.LnLik <- nlm(LnLik, init, hessian = TRUE)
coef.tobitNLM <- est.LnLik$estimate
se.tobitNLM <- sqrt(diag(solve(est.LnLik$hessian)))
```

Second way is to use the function `vglm` in the library `VGAM`. First, we need to declare the tobit distribution (`tobit`), using the `family` augment. The `tobit` function needs the censored point (the value of a) in arguments `Lower` and `Upper`. When you specify `Lower`, the observed outcome is left-censored. On the other hand, when you specify `Upper`, the observed outcome is right-censored. In this application, we set `Lower = 0`.

```r
library(VGAM)
model <- whrs ~ kl6 + k618 + wa + we
tobitVGAM <- vglm(model, family = tobit(Lower = 0), data = dt)
coef.tobitVGAM <- coef(tobitVGAM)
coef.tobitVGAM[2] <- exp(coef.tobitVGAM[2])
se.tobitVGAM <- sqrt(diag(vcov(tobitVGAM)))[-2]
```

2.3 Interpretations

Table 2 shows results of tobit regression estimated by two methods. As a comparison, we also show the OLS result in column (3). Although all specifications show the same sign of coefficients, size of coefficients of censored regression becomes stronger than of OLSE. As with the truncated regression, the number of preschool and school-aged children reduces the hours of work. Unlike the truncated regression, the relationship between married women's characteristics and labor participation is statistically significant. For example, high educated women increases labor time.

```r
ols <- lm(whrs ~ kl6 + k618 + wa + we, data =dt)
names(coef.tobitNLM) <- c("sigma", names(coef(ols)))
```

```r
names(se.tobitNLM) <- c("sigma", names(coef(ols)))
names(coef.tobitVGAM) <- c(names(coef(ols))[1], "sigma", names(coef(ols))[-1])
names(se.tobitVGAM) <- names(coef(ols))

stargazer(
  ols, ols, ols,
  column.labels = c("Tobit (vglm)", "Tobit (nlm)", "OLS"),
  coef = list(coef.tobitVGAM[-2], coef.tobitNLM[-1]),
  se = list(se.tobitVGAM, se.tobitNLM[-1]),
  report = "vcs", keep.stat = c("n"),
  covariate.labels = c(
    "\\#.Preschool Children",
    "\\#.School-aged Children",
    "Age", "Education Years"
  ),
  add.lines = list(
    c("Estimated Sigma",
      round(coef.tobitVGAM[2], 3), round(coef.tobitNLM[1], 3)),
    c("Log-Likelihood",
      round(logLik(tobitVGAM), 3), round(-est.LnLik$minimum, 3))
  ),
  omit.table.layout = "n", table.placement = "t",
  title = "Tobit Regression: Labor Market Participation of Married Women",
  label = "lfp_tobit",
  type = "latex", header = FALSE
)
```

3 Empirical Application of Poisson Regression: Demand of Recreation

3.1 Background and Data

The Poisson distribution is used for drawing purchasing behavior. Especially, the parameter λ means that preference for goods because the expectation of frequency of purchasing, $E(X)$, is equal to λ (we omit proof here). For example, Tsuyoshi Morioka, a famous marketer contributing the v-shaped recovery of Universal Studio Japan, insists that marketers try to increase the parameter λ.

In this application, using cross-section data about recreational boating trips to Lake Somerville, Texas, in 1980, we investigates who has a high preference for this area. We use the built-in dataset called `RecreationDemand` in the library `AER`. This dataset is based on a survey administered to 2,000 registered leisure boat owners in 23 counties in eastern Texas. We use following four variables:

Table 2: Tobit Regression: Labor Market Participation of Married Women

	Dependent variable:		
	whrs		
	Tobit (vglm)	Tobit (nlm)	OLS
	(1)	(2)	(3)
#.Preschool Children	−827.768	−827.733	−462.123
	(218.507)	(171.275)	(124.677)
#.School-aged Children	−140.017	−140.004	−91.141
	(75.203)	(69.379)	(45.850)
Age	−24.980	−24.973	−13.158
	(13.217)	(12.528)	(8.335)
Education Years	103.694	103.707	53.262
	(41.433)	(41.780)	(26.094)
Constant	588.961	588.488	940.059
	(838.808)	(812.625)	(530.720)
Estimated Sigma	1309.928	1309.914	
Log-Likelihood	-1367.09	-1367.09	
Observations	250	250	250

- `trips`: Number of recreational boating trips.
- `income`: Annual household income of the respondent (in 1,000 USD).
- `ski`: Dummy variable taking 1 if the individual was engaged in water-skiing at the lake
- `userfee`: Dummy variable taking 1 if the individual payed an annual user fee at Lake Somerville?

```r
library(AER)
data("RecreationDemand")
summary(RecreationDemand)
```

```
##      trips           quality        ski          income        userfee
##  Min.   : 0.000   Min.   :0.000   no :417   Min.   :1.000   no :646
##  1st Qu.: 0.000   1st Qu.:0.000   yes:242   1st Qu.:3.000   yes: 13
##  Median : 0.000   Median :0.000             Median :3.000
##  Mean   : 2.244   Mean   :1.419             Mean   :3.853
##  3rd Qu.: 2.000   3rd Qu.:3.000             3rd Qu.:5.000
```

```
##  Max.   :88.000   Max.   :5.000            Max.   :9.000
##      costC            costS             costH
##  Min.   :  4.34   Min.   :  4.767   Min.   :  5.70
##  1st Qu.: 28.24   1st Qu.: 33.312   1st Qu.: 28.96
##  Median : 41.19   Median : 47.000   Median : 42.38
##  Mean   : 55.42   Mean   : 59.928   Mean   : 55.99
##  3rd Qu.: 69.67   3rd Qu.: 72.573   3rd Qu.: 68.56
##  Max.   :493.77   Max.   :491.547   Max.   :491.05
```

3.2 Model

Let y_i be the number of recreational boating trips. We assume that this variable follows the Poisson distribution conditional co covariates $\mathbf{x}_i$. That is,

$$p_\beta(y_i|\mathbf{x}_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!},$$

where $\lambda_i = \exp(\mathbf{x}_i\beta)$. Importantly, λ_i represents the preference for boating trips because

$$E[y_i|\mathbf{x}_i] = \lambda_i = \exp(\mathbf{x}_i\beta).$$

Assuming iid sample, the joint density function is defined by

$$p_\beta((y_i|\mathbf{x}_i), i = 1, \dots, n) = \prod_{i=1}^n \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}.$$

Thus, the log-likelihood function is

$$M_n(\beta) = \sum_{i=1}^n (-\lambda_i + y_i \log \lambda_i - \log y_i!) = \sum_{i=1}^n (-\exp(\mathbf{x}_i\beta) + y_i\mathbf{x}_i\beta - \log y_i!).$$

Since the first-order condition (orthogonality condition) is non-linear with respect to β, we apply the Newton-Raphson method to obtain MLE. In `R`, there are two way to implement the Poisson regression. First way is to define the log-likelihood function directly and minimize its function by `nlm` function. We need to give intial values in argument of this function. To set initial values, we assume that coefficients of explanatory variables are zero. Then, we have $E[y_i|\mathbf{x}_i] = \exp(\beta_1) = E[y_i]$ where β_1 is intercept of regression equation. Thus, the initial value of β_1, `b[1]` is $\log E[y_i]$. We replace the expectation of y_i by the mathematical mean of y_i.

```
trips <- RecreationDemand$trips; income <- RecreationDemand$income
ski <- as.integer(RecreationDemand$ski) - 1
userfee <- as.integer(RecreationDemand$userfee) - 1

LnLik <- function(b) {
```

```r
  xb <- b[1] + b[2]*income + b[3]*ski + b[4]*userfee
  LL_i <- -exp(xb) + trips*xb - log(gamma(trips+1))
  LL <- -sum(LL_i)
  return(LL)
}

init <- c(log(mean(trips)), 0, 0, 0)
poissonMLE <- nlm(LnLik, init, hessian = TRUE)
coef.poissonMLE <- poissonMLE$estimate
se.poissonMLE <- sqrt(diag(solve(poissonMLE$hessian)))
logLik.poissonMLE <- -poissonMLE$minimum
```

The second way is to use `glm` function. To implement this function, we need to specify the Poisson distribution, `poisson()` in the `family` augment. We can obtain the value of log-likelihood function, using the `logLik` function.

```r
model <- trips ~ income + ski + userfee
poissonGLM <- glm(model, family = poisson(), data = RecreationDemand)
logLik.poissonGLM <- as.numeric(logLik(poissonGLM))
```

3.3 Interpretations

Table 3 shows results of the Poisson regression estimated by two methods, `nlm` and `glm`. As a comparison, we also show the result of OLS estimation. Clearly, the `nlm` methods (column 1) returns quite similar results to the `glm` method (column 2). Alotough the size of OLSE is farther away from zero than coefficients of the Poisson regression, the sign of OLSE is same as coefficients of the Poisson regression. Surprisingly, we obtain the negative relationship between annual income and preference for boating trips. This implies that high-earners are less likely to go to Lake Somerville.

```r
names(coef.poissonMLE) <- names(coef(poissonGLM))
names(se.poissonMLE) <- names(coef(poissonGLM))
ols <- lm(model, data = RecreationDemand)

stargazer(
  poissonGLM, poissonGLM, ols,
  coef = list(coef.poissonMLE),
  se = list(se.poissonMLE),
  report = "vcs", keep.stat = c("n"),
  covariate.labels = c(
    "Income",
    "1 = Playing water-skiing",
    "1 = Paying annual fee"
  ),
  add.lines = list(
```

Table 3: Poisson Regression: Recreation Demand

	Dependent variable:		
	trips		
	Poisson		*OLS*
	(1)	(2)	(3)
Income	−0.146	−0.146	−0.277
	(0.017)	(0.017)	(0.133)
1 = Playing water-skiing	0.547	0.547	1.243
	(0.055)	(0.055)	(0.509)
1 = Paying annual fee	1.904	1.904	12.412
	(0.078)	(0.078)	(1.688)
Constant	1.006	1.006	2.609
	(0.065)	(0.065)	(0.545)
Method	nlm	glm	
Log-Likelihood	-2529.256	-2529.256	
Observations	659	659	659

```
    c("Method", "nlm", "glm", ""),
    c("Log-Likelihood",
      round(logLik.poissonMLE, 3), round(logLik.poissonGLM, 3), "")
  ),
  omit.table.layout = "n", table.placement = "t",
  title = "Poisson Regression: Recreation Demand",
  label = "recreation",
  type = "latex", header = FALSE
)
```