

# 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.<sup>1</sup> This dataset contains the following variables:

- **whrs**: Hours of work. This outcome variable is truncated from below at zero.
- **k16**: 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           k16           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
```

---

<sup>1</sup><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 participate in the labor market (`whrs = 0`), we use the truncated regression model. Thus, the selection rule is as follows:

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

where  $u_i \sim N(0, \sigma^2)$ . In this case,  $a_1 = 0$  and  $a_2 = +\infty$ .

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

$$p_\theta(y_i|x_i, s_i = 1) = \frac{f(y_i|x_i)}{\int_0^{+\infty} f(y_i|x_i)dy_i}.$$

where  $\theta = (\beta, \sigma^2)'$ . Because the distribution of  $y_i$  depends on the distribution of  $u_i$ , using  $u_i = y_i - x_i\beta$ , we obtain

$$p_\theta(u_i|x_i, -x_i\beta < u_i) = \frac{1}{\sigma} \frac{\phi(\frac{y_i - x_i\beta}{\sigma})}{1 - \Phi(\frac{-x_i\beta}{\sigma})}.$$

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 initial values in argument of this function. Coefficients of explanatory variables, `b[3:6]`, are zero, and intercept, `b[2]`, and  $\sigma$ , `b[1]`, are given by mean and standard deviation of `whrs`, respectively.

```
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`. This function must specify the truncated point in 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 apply the truncated regression. Although the relationship between labor participation and women's characteristics is statistically insignificant, size of coefficients largely differs among three specifications.

```

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)))
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"
  )
)

```

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

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

```

),
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. Our dataset comes from Stata sample data, which is same as the previous dataset. Unlike the previous dataset, we now observe married women who do not participate in the labor market. 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 also use observations with `lfp = 0` to estimate tobit model.

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

```
##      lfp      whrs      kl6      k618      wa
## 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. This is because married women chooses zero labor time if, without any constraint, the optimal labor time is negative (corner solution). 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 returning 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 standard normal distribution. Note that the last equality 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 joint 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 initial values in argument of this function. As initial values, coefficients of explanatory variables, `b[3:6]`, are zero, and intercept, denoted by `b[2]`, and  $\sigma$ , denoted by `b[1]`, are given by mean and standard deviation of `whrs`, respectively.

```
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`.

```
library(VGAM)
tobitVGAM <- vglm(whrs ~ kl6 + k618 + wa + we, 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 OLS. For example, 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 apply the censored regression.

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

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

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

```

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  $\lambda$  means that preference for goods because the expectation of frequency of purchasing,  $E(X)$ , is equal to  $\lambda$  (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  $\lambda$ .

In this application, using cross-section data about recreational boating trips to Lake Somerville, Texas, in 1980, we investigate 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:

- `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 paid an annual user fee at Lake Somerville?

```
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,  $\lambda_i$  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  $\beta$ , 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. As intial values, coefficients of explanatory variables, `b[2:4]`, are zero, and intercept, denoted by `b[2]`, are given by mean of `trips`.

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

LnLik <- function(b) {
  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(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 argument. We can obtain the value of log-likelihood function, using the `logLik` function.

```
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). Although the size of coefficients of OLSE is farther away from zero than coefficients of the Poisson regression, the sign of coefficients 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.

```
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(
    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
)
```

Table 3: Poisson Regression: Recreation Demand

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