Econometrics II TA Session *

Wang Xin

Nov 20, 2023

For next week, the TA session is on 29th (Wednesday) 10:30, not on 27th (Monday).
The Econometrics II will be carried both on 28th (Tuesday) and on 30th (Thursday).
From December, the TA session will be given at 15:10 every Monday.

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 data set of married women, we explore how much childcare prevents married women to participate in labor market.

Our data set originally comes from Stata sample data [1]. This data set 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

*The codes are cited from documents by Hiroki Kato.
[1] https://www.stata-press.com/data/r18/laborsub.dta

```
library(haven)
dt <- read_dta(file = "./laborsub.dta")
dt <- dt[dt$lfp != 0, c("whrs","kl6","k618","wa","we")]
summary(dt)
```

```
      whrs            kl6              k618
 Min.   :  12   Min.   :0.0000   Min.   :0.000
 1st Qu.: 645   1st Qu.:0.0000   1st Qu.:0.000
 Median :1406   Median :0.0000   Median :1.000
 Mean   :1333   Mean   :0.1733   Mean   :1.313
 3rd Qu.:1903   3rd Qu.:0.0000   3rd Qu.:2.000
 Max.   :4950   Max.   :2.0000   Max.   :8.000
       wa              we
 Min.   :30.00   Min.   : 6.00
 1st Qu.:35.00   1st Qu.:12.00
 Median :43.50   Median :12.00
 Mean   :42.79   Mean   :12.64
 3rd Qu.:48.75   3rd Qu.:13.75
 Max.   :60.00   Max.   :17.00
```

1.2 Model

As have reviewed, the truncated regression model is given as

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + u_i \quad \text{if } a_1 < y < a_2$$

where $u_i \sim N(0, \sigma^2)$, and the probability density function of y_i is

$$p_\theta(y_i|\mathbf{x}_i, a_1 < y < a_2) = \frac{f(y_i|\mathbf{x}_i)}{P(a_1 < y < a_2|\mathbf{x}_i)}$$
$$= \frac{\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}{\Phi\left(\frac{a_2 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{a_1 - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)}$$

We still introduce MLE to estimate this model. Remind that the criterion function is

$$M_n(\theta) = \sum_{i=1}^{n} \log p_\theta(y_i|\mathbf{x}_i, a_1 < y < a_2).$$

In this case, we only observe the women participated in work. Thus, the selection rule is as follows:

$$\texttt{whrs} = \beta_1 + \beta_2\texttt{kl6} + \beta_3\texttt{k618} + \beta_4\texttt{wa} + \beta_5\texttt{we} + u_i \quad \text{when } \texttt{whrs} > 0$$

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. 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 is the standard deviation of `whrs`, and we denote it as `b[1]`, and the initial value of β_1 is the mean of `whrs`, denoted as `b[2]`. Note that these initial values are not unbiased estimator.

```
# Newton method
whrs <- dt$whrs
kl6 <- dt$kl6; k618 <- dt$k618
wa <- dt$wa; we <- dt$we
LnLik <- function(b) {
  sigma <- b[1]
  xb <- b[2] + b[3]*kl6 + b[4]*k618 + b[5]*wa + b[6]*we
  condp <- dnorm((whrs - xb)/sigma)/(1 - pnorm(-xb/sigma))
  # dnorm() generate a normal probability density function
  LL_i <- log(condp/sigma)
  LL <- -sum(LL_i)
  return(LL)
}
init <- c(sd(whrs), mean(whrs), 0, 0, 0, 0)
est.LnLik <- nlm(LnLik, init, hessian = TRUE)
coef.LnLik <- est.LnLik$estimate
se.LnLik <- sqrt(diag(solve(est.LnLik$hessian)))
```

Second way is to use the function `truncreg` in the library `truncreg`. We must specify the truncated point, using `point` and `direction` arguments. The point argument indicates where the outcome variable is truncated. 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`.

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

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

	Truncated (truncreg)	Truncated (nlm)	OLS
#.Preschool Children	−803.004**	−803.032***	−421.482**
	(321.361)	(252.803)	(167.973)
#.School-aged Children	−172.875*	−172.875*	−104.457*
	(88.729)	(100.590)	(54.186)
Age	−8.821	−8.821	−4.785
	(14.368)	(14.646)	(9.691)
Education Years	16.529	16.529	9.353
	(46.504)	(46.430)	(31.238)
Constant	1,586.260*	1,586.228*	1,629.817***
	(912.354)	(932.878)	(615.130)
Estimated Sigma	983.726	983.736	
Log-Likelihood	-1200.916	-1200.916	
Observations	150	150	150

1.3 Interpretation and Model Fitness

Table 1 shows results of truncated regression estimated by two methods. As a comparison, we also show the OLS result in column (3).

```
ols <- lm(model, data = dt)
```

All specifications show that the number of preschool and school-aged children reduce the hours of work. The magnitude 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`.

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

2.1 Background and Data

We continue to investigate the previous research question. We use data set coming from same source as the previous one. Unlike the previous data set, we now observe

married women 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 data set contains observations with `lfp = 1`. In this application, we use observations with `lfp = 0` to estimate the tobit model.

```
library(haven)
dt <- read_dta(file = "./laborsub.dta")
summary(dt)
      lfp             whrs              kl6
 Min.   :0.0   Min.   :   0.0   Min.   :0.000
 1st Qu.:0.0   1st Qu.:   0.0   1st Qu.:0.000
 Median :1.0   Median : 406.5   Median :0.000
 Mean   :0.6   Mean   : 799.8   Mean   :0.236
 3rd Qu.:1.0   3rd Qu.:1599.8   3rd Qu.:0.000
 Max.   :1.0   Max.   :4950.0   Max.   :3.000
      k618              wa              we
 Min.   :0.000   Min.   :30.00   Min.   : 5.00
 1st Qu.:0.000   1st Qu.:35.00   1st Qu.:12.00
 Median :1.000   Median :43.00   Median :12.00
 Mean   :1.364   Mean   :42.92   Mean   :12.35
 3rd Qu.:2.000   3rd Qu.:49.00   3rd Qu.:13.00
 Max.   :8.000   Max.   :60.00   Max.   :17.00
```

2.2 Model

When the dependent variable is censored, values in a certain range are all transformed to (or reported as) a single value. Conventional regression methods fail to account for the qualitative difference between limit observations and nonlimit (continuous) observations.

In this case, we should use the tobit model, which is

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

where u_i is usually assumed as in $N(0, \sigma^2)$.

The probability function of the censored y_i is defined by

$$p_{\beta,\sigma^2}(y_i|\mathbf{x}_i) = F_y(a)^{I_{y_i=a}} f(y_i|\mathbf{x}_i)^{1-I_{y_i=a}}$$

where $f(y_i|\mathbf{x}_i)$ denotes the probability of y_i in the whole interval conditionally on $\mathbf{x}_i$. $1[y_i = 0]$ is an indicator function returing 1 if $y_i = 0$.

In this case, $a = 0$, and the probability that y_i is not greater than 0 should be:

$$F(y_i \leq 0) = F(u_i \geq \mathbf{x}_i\boldsymbol{\beta}) = 1 - \Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)$$

Same as the probability density function we have derived in last week, that

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

and the criterion function under iid assumption is derived as

$$M_n(\beta, \sigma^2) = \log \prod_{i=1}^{n} \left(1 - \Phi\left(\frac{\mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right)^{I_{y_i=0}} \left(\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i\boldsymbol{\beta}}{\sigma}\right)\right)^{1-I_{y_i=0}}$$

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^2, `b[1]` is the standard deviation of `whrs`, and the initial value of β_1, `b[2]` is the mean of `whrs`.

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

# Newton-Raphson method
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`.

```
# tobitVGAM function
library(VGAM)
```

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

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

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.

```
ols <- lm(whrs ~ kl6 + k618 + wa + we, data =dt)
```

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 data set called `RecreationDemand` in the library `AER`. This data set 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 payed an annual user fee at Lake Somerville?

```
library(AER)
data(RecreationDemand)
dt <- RecreationDemand
dt <- dt[,c("trips", "ski", "income", "userfee")]
summary(dt)
     trips          ski           income      userfee
 Min.   : 0.000   no :417   Min.   :1.000   no :646
 1st Qu.: 0.000   yes:242   1st Qu.:3.000   yes: 13
 Median : 0.000             Median :3.000
 Mean   : 2.244             Mean   :3.853
 3rd Qu.: 2.000             3rd Qu.:5.000
 Max.   :88.000             Max.   :9.000
```

3.2 Model

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

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

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

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

Assuming iid sample, the log-likelihood function is

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

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 <- dt$trips
income <- dt$income
ski <- as.integer(dt$ski) - 1
userfee <- as.integer(dt$userfee) - 1

# nlm()
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(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.

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

Table 3: Poisson Regression: Recreation Demand

	Dependent variable:	
	trips	
	(1)	(2)
Income	-0.146^{***}	-0.146^{***}
	(0.017)	(0.017)
1 = Playing water-skiing	0.547^{***}	0.547^{***}
	(0.055)	(0.055)
1 = Paying annual fee	1.904^{***}	1.904^{***}
	(0.078)	(0.078)
Constant	1.006^{***}	1.006^{***}
	(0.065)	(0.065)
Method	nlm	glm
Log-Likelihood	-2529.256	-2529.256
Observations	659	659

3.3 Interpretations

Table 3 shows results of the Poisson regression estimated by two methods, `nlm` and `glm`. Clearly, the `nlm` methods (column 1) returns quite similar results to the `glm` method (column 2). 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.