Econometrics II TA Session #3

Hiroki Kato

1 Empirical Application of Binary Model: Racial Discrimination in Court

Brief Background. Recently, in the U.S., anti-racism activities called "Black Lives Matter" are getting hot. These activities stems from the death of George Floyd, who was killed by a white police officer on May 25, 2020. The empirical application of binary model investigates whether the judgement of death penalty is based on race of defendant and race of victim.

Data. The package `catdata` contains many built-in dataets which include categorical variables. We use the built-in dataset `deathpenalty` which is about the death-penalty judgement of defendants in cases of multiple murders in Florida between 1976 and 1987.

```
dt <- read.csv(
  file = "./data/titanic.csv",
  header = TRUE,  sep = ",", row.names = NULL,  stringsAsFactors = FALSE)
dt <- dt[,c("survived", "age", "fare", "sex")]
head(dt)
```

```
##   survived   age     fare    sex
## 1        1 29.00 211.3375 female
## 2        1  0.92 151.5500   male
## 3        0  2.00 151.5500 female
## 4        0 30.00 151.5500   male
## 5        0 25.00 151.5500 female
## 6        1 48.00  26.5500   male
```

This dataset contains three dummy variables

1. `DeathPenalty` is a dummy variable taking 1 if the judgement is death penalty.
2. `VictimRace` is a dummy variable taking 1 if the race of the victim is white.
3. `DefendantRace` is a dummy variable taking 1 if the race of the defendant is white.

This dataset aggregates observations with repect to `DeathPenalty`, `VictimRace` and `DefendantRace`. The variable `Freq` represents the number of observations. Since it is inconvinient for us to use the original data for estimation, we disaggregate this dataset. For example, we make 37 rows whose elements are `DeathPenalty = 0`, `VictimRace = 1`, and `DefendantRace = 0` because there are 37 obaservations, i.e., `Freq = 37`.

```r
dt <- subset(dt, !is.na(survived)&!is.na(age)&!is.na(fare)&!is.na(sex))
dt$female <- ifelse(dt$sex == "female", 1, 0)
```

Model. In a binary model, a dependent (outcome) variable y_i takes only two values, i.e., $y_i \in \{0,1\}$. A binary variable is sometimes called a *dummy* variable. In this application, the outcome variable is `DeathPenalty` taking 1 if the judgement is death penalty. We make three explanatory variables.

1. `WB` is a dummy variable taking 1 if the race of the victim and the defendant is white and black, respectively.
2. `BW` is a dummy variable taking 1 if the race of the victim and the defendant is black and white, respectively.
3. `WW` is a dummy variable taking 1 if the race of both the victim and the defendant is black.

The regression function is

$$
\begin{aligned}
&\mathbb{E}[DeathPenalty|WB, BW, WW] \\
=&\mathbb{P}[DeathPenalty = 1|WB, BW, WW] = G(\beta_0 + \beta_1 WB + \beta_2 BW + \beta 3WW).
\end{aligned}
$$

The function $G(\cdot)$ is arbitrary function. In practice, we often use following three specifications:

- Linear probability model (LPM): $G(\mathbf{x}_i\beta) = \mathbf{x}_i\beta$.
- Probit model: $G(\mathbf{x}_i\beta) = \Phi(\mathbf{x}_i\beta)$ where $\Phi(\cdot)$ is the standard Gaussian cumulative function.
- Logit model: $G(\mathbf{x}_i\beta) = 1/(1+\exp(-\mathbf{x}_i\beta))$.

1.1 Linear Probability Model

The linear probability model is

$$
\mathbb{P}[DeathPenalty = 1|WB, BW, WW] = \beta_0 + \beta_1 WB + \beta_2 BW + \beta 3WW
$$

This model can be estimated using the OLS method. In `R`, we can use the OLS method, running `lm()` function.

```r
model <- survived ~ female + age + fare
LPM <- lm(model, data = dt)
```

However, `lm()` function does not deal with heteroskedasticity problem. To resolve it, we need to claculate heteroskedasticity-robust standard errors using the White method.

$$
\hat{V}(\hat{\beta}) = \left(\frac{1}{n}\sum_i \mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\frac{1}{n}\sum_i \hat{u}_i^2\mathbf{x}_i'\mathbf{x}_i\right)\left(\frac{1}{n}\sum_i \mathbf{x}_i'\mathbf{x}_i\right)^{-1}
$$

```r
# heteroskedasticity-robust standard errors
dt$"(Intercept)" <- 1
X <- as.matrix(dt[,c("(Intercept)", "female", "age", "fare")])
u <- diag(LPM$residuals^2)

XX <- t(X) %*% X
avgXX <- XX * nrow(X)^{-1}
inv_avgXX <- solve(avgXX)

uXX <- t(X) %*% u %*% X
avguXX <- uXX * nrow(X)^{-1}

vcov_b <- (inv_avgXX %*% avguXX %*% inv_avgXX) * nrow(X)^{-1}
rse_b <- sqrt(diag(vcov_b))

# homoskedasticity-based standard errors
se_b <- sqrt(diag(vcov(LPM)))

print("The Variance of OLS"); vcov(LPM)
```

```
## [1] "The Variance of OLS"

##               (Intercept)        female           age          fare
## (Intercept)  9.754357e-04 -2.891381e-04 -2.333963e-05 -3.329763e-07
## female      -2.891381e-04  7.136865e-04  2.373259e-06 -1.272800e-06
## age         -2.333963e-05  2.373259e-06  8.026024e-07 -4.090649e-08
## fare        -3.329763e-07 -1.272800e-06 -4.090649e-08  5.524412e-08
```

```r
print("The Robust variance of OLS"); vcov_b
```

```
## [1] "The Robust variance of OLS"

##               (Intercept)        female           age          fare
## (Intercept)  1.133289e-03 -2.798532e-04 -2.789675e-05  2.813843e-07
## female      -2.798532e-04  7.903766e-04  3.169092e-06 -2.401923e-06
## age         -2.789675e-05  3.169092e-06  8.857523e-07 -3.650375e-08
## fare         2.813843e-07 -2.401923e-06 -3.650375e-08  4.071639e-08
```

```r
print("The Robust se using White method"); rse_b
```

```
## [1] "The Robust se using White method"

##  (Intercept)       female          age         fare
## 0.0336643606 0.0281136372 0.0009411442 0.0002017830
```

```r
print("The Robust t-value using White method"); coef(LPM)/rse_b
```

```
## [1] "The Robust t-value using White method"
```

```
## (Intercept)      female         age        fare 
##    6.482874   18.229508   -1.884168    7.162302
```

Using the package `lmtest` and `sandwich` is the most easiest way to calculate heteroskedasticity-robust standard errors and t-statistics.

```r
library(lmtest) #use function `coeftest`
library(sandwich) #use function `vcovHC`
coeftest(LPM, vcov = vcovHC(LPM, type = "HC0"))[, "Std. Error"]
```

```
##  (Intercept)       female          age         fare 
## 0.0336643606 0.0281136372 0.0009411442 0.0002017830
```

```r
coeftest(LPM, vcov = vcovHC(LPM, type = "HC0"))[, "t value"]
```

```
## (Intercept)      female         age        fare 
##    6.482874   18.229508   -1.884168    7.162302
```

Finally, we obtain follwing results of linear probability model. We will discuss interpretation of results and goodness-of-fit of LPM later.

```r
# t-stats
t_b <- coef(LPM)/se_b
rt_b <- coef(LPM)/rse_b
# p-value Pr( > |t|)
p_b <- pt(abs(t_b), df = nrow(X)-ncol(X), lower = FALSE)*2
rp_b <- pt(abs(rt_b), df = nrow(X)-ncol(X), lower = FALSE)*2

library(stargazer)
stargazer(
  LPM, LPM,
  se = list(se_b, rse_b), t = list(t_b, rt_b), p = list(p_b, rp_b),
  t.auto = FALSE, p.auto = FALSE,
  report = "vcstp", keep.stat = c("n"),
  add.lines = list(
    c("Standard errors", "Homoskedasticity-based", "Heteroskedasticity-robust")),
  title = "Results of Linear Probability Model",
  type = "latex", header = FALSE, font.size = "small",
  omit.table.layout = "n"
)
```

Table 1: Results of Linear Probability Model

	Dependent variable:	
	survived	
	(1)	(2)
female	0.512	0.512
	(0.027)	(0.028)
	t = 19.184	t = 18.230
	p = 0.000	p = 0.000
age	−0.002	−0.002
	(0.001)	(0.001)
	t = −1.979	t = −1.884
	p = 0.049	p = 0.060
fare	0.001	0.001
	(0.0002)	(0.0002)
	t = 6.149	t = 7.162
	p = 0.000	p = 0.000
Constant	0.218	0.218
	(0.031)	(0.034)
	t = 6.988	t = 6.483
	p = 0.000	p = 0.000
Standard errors	Homoskedasticity-based	Heteroskedasticity-robust
Observations	1,045	1,045