TA Session 3

Chi-Yuan Fang

March 16, 2021

Contents

1 Introduction

1.1 TA Information

TA: Chi-Yuan Fang

TA sessions: Tuesday 1:20 – 3:10 PM (SS 501)

Email: r09323017@ntu.edu.tw

Office hours: Friday 2:00 – 3:30 PM or by appointments (SS 643)

Class group on Facebook: Statistics (Fall 2020) and Econometrics (Spring 2021)

https://www.facebook.com/groups/452292659024369/

Because screens are not clear in SS 501, I will provide the link of live streaming in the group.

1.2 TA Sessions Schedule

Week	TA Sessions	Quiz	Content	Remind
1	02/23: No class			
2	03/02: Class 1		Function, Confidence Interval, T test	03/10 Turn in HW1
3	03/09: Class 2		Loops, Linear Model	03/10 Turn in HW1, 03/16 Quiz 1
4	03/16: Class 3	Quiz 1	OLS	03/24 Turn in HW2
5	03/23: Class 4		Multiple Regression	03/24 Turn in HW2, 03/30 Quiz 2
6	03/30: Class 5	Quiz 2		04/14 Turn in HW3
7	04/06: No class			04/14 Turn in HW3
8	04/13: Class 6			04/14 Turn in HW3, 04/20 Quiz 3
9	04/20: Class 7	Quiz 3		**04/28 Midterm**
10	04/27: Class 8		Review and Q&A	**04/28 Midterm**, 05/05 Turn in HW4
11	05/04: Class 9			05/05 Turn in HW4, 05/11 Quiz 4
12	05/11: Class 10	Quiz 4		05/19 Turn in HW5

Week	TA Sessions	Quiz	Content	Remind
13	05/18: Class 11			05/19 Turn in HW5, 05/25 Quiz 5
14	05/25: Class 12	Quiz 5		06/02 Turn in HW6
15	06/01: Class 13			06/02 Turn in HW6, 06/08 Quiz 6
16	06/08: Class 14	Quiz 6	Review and Q&A	**06/16 Final Exam**
17	06/15: No class			**06/16 Final Exam**
18	06/22: No class			

1.3 Reference

Introduction to Econometrics with R

https://www.econometrics-with-r.org

R for Data Science

https://r4ds.had.co.nz

R Markdown

https://rmarkdown.rstudio.com

Introduction to R Markdown

https://rpubs.com/brandonkopp/RMarkdown

What is a good book on learning R with examples?

https://www.quora.com/What-is-a-good-book-on-learning-R-with-examples

2 Empirical Exercise 6.1

Use the **Birthweight_Smoking** data set introduced in Empirical Exercise E5.3 to answer the following questions.

a. Regress *Birthweight* on *Smoker*. What is the estimated effect of smoking on birth weight?

Solution

```
library(ggplot2); library(dplyr);
library(jtools); library(ggstance);
```

```
##
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
```

```
library(broom.mixed); library(huxtable)
```

```
## Warning in checkMatrixPackageVersion(): Package version inconsistency detected.
## TMB was built with Matrix version 1.3.2
## Current Matrix version is 1.2.18
## Please re-install 'TMB' from source using install.packages('TMB', type = 'source') or ask CRAN for a

## Registered S3 method overwritten by 'broom.mixed':
##   method      from
##   tidy.gamlss broom
```

```
##
## Attaching package: 'huxtable'

## The following object is masked from 'package:dplyr':
##
##     add_rownames

## The following object is masked from 'package:ggplot2':
##
##     theme_grey
```

```r
# import data
library(readxl)
Birthweight_Smoking <- read_excel("Birthweight_Smoking/Birthweight_Smoking.xlsx")

E61a <- lm(birthweight ~ smoker, data = Birthweight_Smoking)

summ(E61a, confint = TRUE, digits = 4)
```

```
## MODEL INFO:
## Observations: 3000
## Dependent Variable: birthweight
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,2998) = 88.2793, p = 0.0000
## R² = 0.0286
## Adj. R² = 0.0283
##
## Standard errors: OLS
## ------------------------------------------------------------------------
##                           Est.        2.5%       97.5%     t val.        p
## ----------------- ----------- ----------- ----------- ---------- --------
## (Intercept)          3432.0600   3408.7840   3455.3359   289.1154   0.0000
## smoker               -253.2284   -306.0736   -200.3831    -9.3957   0.0000
## ------------------------------------------------------------------------
```

b. Regress *Birthweight* on *Smoker*, *Alcohol*, and *Nprevist*.
   i. Using the two conditions in Key Concept 6.1, explain why the exclusion of *Alcohol* and *Nprevist* could lead to omitted variable bias in the regression estimated in (a).
   ii. Is the estimated effect of smoking on birth weight substantially different from the regression that excludes *Alcohol* and *Nprevist*? Does the regression in (a) seem to suffer from omitted variable bias?
   iii. Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.
   iv. Compute R^2 and $\overline{R}^2$. Why are they so similar?
   v. How should you interpret the coefficient on Nprevist? Does the coefficient measure a causal effect of prenatal visits on birth weight? If not, what does it measure?

Solution

```r
# nprevist: total number of prenatal visits
E61b <- lm(birthweight ~ smoker + alcohol + nprevist, data = Birthweight_Smoking)

summ(E61b, confint = TRUE, digits = 4)
```

```
## MODEL INFO:
```

```
## Observations: 3000
## Dependent Variable: birthweight
## Type: OLS linear regression
##
## MODEL FIT:
## F(3,2996) = 78.4697, p = 0.0000
## R² = 0.0729
## Adj. R² = 0.0719
##
## Standard errors: OLS
## ------------------------------------------------------------------------
##                           Est.        2.5%       97.5%    t val.        p
## ----------------- ----------- ----------- ----------- ---------- --------
## (Intercept)          3051.2486   2984.5516   3117.9456    89.7005   0.0000
## smoker               -217.5801   -269.8923   -165.2679    -8.1553   0.0000
## alcohol               -30.4913   -179.9677    118.9851    -0.4000   0.6892
## nprevist               34.0699     28.4720     39.6679    11.9334   0.0000
## ------------------------------------------------------------------------
```

i.

- Smoking may be correlated with both alcohol and the number of pre-natal doctor visits.
- Both alcohol consumption and the number of doctor visits may have their own independent affects on birthweight.

ii. Because the estimated is smaller than (a), the regression in (a) may suffer from omitted variable bias.

```r
export_summs(E61a, E61b,
             model.names = c("Model (a)", "Model (b)"))
```

	Model (a)	Model (b)
(Intercept)	3432.06 ***	3051.25 ***
	(11.87)	(34.02)
smoker	-253.23 ***	-217.58 ***
	(26.95)	(26.68)
alcohol		-30.49
		(76.23)
nprevist		34.07 ***
		(2.85)
N	3000	3000
R2	0.03	0.07

*** p < 0.001; ** p < 0.01; * p < 0.05.

iii.

```r
# predict value
E61biii <- function(x){
  E61b$coefficients %*% matrix(c(1, x), ncol = 1)

}

# predict value:
# smoker = 1, alcohol = 0, nprevist = 8
E61biii(c(1, 0, 8))
```

```
##          [,1]
## [1,] 3106.228
```

iv. Because the sample size is very large, they are nearly identical.

v. *Nprevist* is a control variable. It captures mother's access to healthcare and health. Thus, its coefficient does not have a causal interpretation.

c. Estimate the coefficient on Smoking for the multiple regression model in (b), using the three-step process in Appendix 6.3 (the Frisch–Waugh theorem). Verify that the three-step process yields the same estimated coefficient for *Smoking* as that obtained in (b).

Solution

The OLS estimator in multiple regression can be computed by a sequence of shorter regressions. Consider the multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, \ldots, n. \tag{1}$$

The OLS estimator of β_1 can be computed in three steps:

1. Regress X_1 on $X_2, X_3, \ldots, X_k$, and let $\widetilde{X}_1$ denote the residuals from this regression;
2. Regress Y on $X_2, X_3, \ldots, X_k$, and let $\widetilde{Y}$ denote the residuals from this regression;
3. Regress $\widetilde{Y}$ on $\widetilde{X}_1$.

```r
# Step 1: regress X1 on X2, X3, ..., Xk, and residuals = X1 tilde
E61c1 <- lm(smoker ~ alcohol + nprevist, data = Birthweight_Smoking)
smoker <- E61c1$residuals

# Step 2: regress Y on X2, X3, ..., Xk, and residuals = Y tilde
E61c2 <- lm(birthweight ~ alcohol + nprevist, data = Birthweight_Smoking)
birthweight <- E61c2$residuals

# Step 3: regress Y tilde on X1 tilde
E61c3 <- lm(birthweight ~ smoker)

summ(E61c3)
```

```
## MODEL INFO:
## Observations: 3000
## Dependent Variable: birthweight
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,2998) = 66.55, p = 0.00
## R² = 0.02
## Adj. R² = 0.02
```

```
##
## Standard errors: OLS
## ---------------------------------------------------
##                           Est.     S.E.    t val.      p
## ----------------- ---------- -------- --------- ------
## (Intercept)               0.00    10.41      0.00   1.00
## smoker                 -217.58    26.67     -8.16   0.00
## ---------------------------------------------------
```

```r
# comparison
export_summs(E61b, E61c3,
             model.names = c("Model (b)", "Model (c)"))
```

	Model (b)	Model (c)
(Intercept)	3051.25 ***	0.00
	(34.02)	(10.41)
smoker	-217.58 ***	-217.58 ***
	(26.68)	(26.67)
alcohol	-30.49	
	(76.23)	
nprevist	34.07 ***	
	(2.85)	
N	3000	3000
R2	0.07	0.02

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

d. An alternative way to control for prenatal visits is to use the binary variables $Tripre0$ through $Tripre3$. Regress $Birthweight$ on $Smoker$, $Alcohol, Tripre0, Tripre2$, and $Tripre3$.
   i. Why is $Tripre1$ excluded from the regression? What would happen if you included it in the regression?
   ii. The estimated coefficient on $Tripre0$ is large and negative. What does this coefficient measure? Interpret its value.
   iii. Interpret the value of the estimated coefficients on $Tripre2$ and $Tripre3$.
   iv. Does the regression in (d) explain a larger fraction of the variance in birth weight than the regression in (b)?

Solution

```r
E61d <- lm(birthweight ~ smoker + alcohol + tripre0 + tripre2 + tripre3, data = Birthweight_Smoking)

summ(E61d, confint = TRUE, digits = 4)
```

```
## MODEL INFO:
## Observations: 3000
## Dependent Variable: birthweight
## Type: OLS linear regression
```

```
##
## MODEL FIT:
## F(5,2994) = 29.1795, p = 0.0000
## R² = 0.0465
## Adj. R² = 0.0449
##
## Standard errors: OLS
## ------------------------------------------------------------------------
##                           Est.        2.5%       97.5%     t val.        p
## ----------------- ----------- ----------- ----------- ---------- --------
## (Intercept)          3454.5493   3429.7449   3479.3538   273.0768   0.0000
## smoker               -228.8476   -282.1112   -175.5840    -8.4244   0.0000
## alcohol               -15.1000   -167.1382    136.9383    -0.1947   0.8456
## tripre0              -697.9687   -907.5260   -488.4114    -6.5307   0.0000
## tripre2              -100.8373   -158.9127    -42.7618    -3.4045   0.0007
## tripre3              -136.9553   -253.7798    -20.1308    -2.2986   0.0216
## ------------------------------------------------------------------------
```

i. $Tripre1$ is omitted to avoid perfect multicollinearity. If we include it in the regression, then coefficient of $Tripre3$ disappears.

```r
E61di <- lm(birthweight ~ smoker + alcohol + tripre0 + tripre1 + tripre2 + tripre3, data = Birthweight_S

summ(E61di, confint = TRUE, digits = 4)
```

```
## MODEL INFO:
## Observations: 3000
## Dependent Variable: birthweight
## Type: OLS linear regression
##
## MODEL FIT:
## F(5,2994) = 29.1795, p = 0.0000
## R² = 0.0465
## Adj. R² = 0.0449
##
## Standard errors: OLS
## -----------------------------------------------------------------------
##                           Est.        2.5%       97.5%    t val.        p
## ----------------- ----------- ----------- ----------- --------- --------
## (Intercept)          3317.5941   3201.9115   3433.2766   56.2314   0.0000
## smoker               -228.8476   -282.1112   -175.5840   -8.4244   0.0000
## alcohol               -15.1000   -167.1382    136.9383   -0.1947   0.8456
## tripre0              -561.0135   -798.0225   -324.0044   -4.6412   0.0000
## tripre1               136.9553     20.1308    253.7798    2.2986   0.0216
## tripre2                36.1180    -89.7106    161.9466    0.5628   0.5736
## tripre3
## -----------------------------------------------------------------------
```

ii. On average, babies born to women who had no prenatal doctor visits ($Tripre0 = 1$) had birthweights thatwere 697.9687 grams lower than babies from others who saw a doctor during the first trimester ($Tripre1 = 1$).

iii.

- On average, babies born to women whose first doctor visit was during the second trimester ($Tripre2 = 1$) had birthweights that were 100.8373 grams lower than babies from others who saw a doctor during the first trimester ($Tripre1 = 1$).

- On average, babies born to women whose first doctor visit was during the third trimester ($Tripre3 = 1$) had birthweights that on average were 136.9553 grams lower than babies from others who saw a doctor during the first trimester ($Tripre1 = 1$).

iv. No, it doesn't. R^2 in (d) is smaller than (b).

```
export_summs(E61b, E61d,
             model.names = c("Model (b)", "Model (d)"))
```

	Model (b)	Model (d)
(Intercept)	3051.25 ***	3454.55 ***
	(34.02)	(12.65)
smoker	-217.58 ***	-228.85 ***
	(26.68)	(27.16)
alcohol	-30.49	-15.10
	(76.23)	(77.54)
nprevist	34.07 ***	
	(2.85)	
tripre0		-697.97 ***
		(106.88)
tripre2		-100.84 ***
		(29.62)
tripre3		-136.96 *
		(59.58)
N	3000	3000
R2	0.07	0.05

*** p < 0.001; ** p < 0.01; * p < 0.05.

3 Empirical Exercise 6.2

Using the data set **Growth** described in Empirical Exercise E4.1, but excluding the data for Malta, carry out the following exercises.

a. Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series $Growth$, $TradeShare$, $YearsSchool$, Oil, Rev_Coups, $Assassinations$, and $RGDP60$. Include the appropriate units for all entries.

Solution

```
# import data
library(readxl)
```

```r
Growth <- read_xlsx("Growth/Growth.xlsx")

E62a <- function(x){
  # meam
  mu <- mean(x)

  # standard deviation
  SD <- sd(x)

  # minimum
  MIN <- min(x)

  # maximum
  MAX <- max(x)

  Table <- data.frame(mu, SD, MIN, MAX)

  colnames(Table) <- c("Mean", "Standard Deviation", "Minimum", "Maximum")

  Table

}

E62a_output <- as.list(matrix(ncol = 4))



for (i in 2:ncol(Growth)){
  E62a_output[[i]] <- E62a(Growth[[i]])

}

names(E62a_output) <- variable.names(Growth)

E62a_output
```

```
## $country_name
## [1] NA
##
## $growth
##       Mean Standard Deviation   Minimum  Maximum
## 1 1.942715            1.89712 -2.811944 7.156855
##
## $oil
##   Mean Standard Deviation Minimum Maximum
## 1    0                  0       0       0
##
## $rgdp60
##       Mean Standard Deviation  Minimum  Maximum
## 1 3103.785           2512.657 366.9999 9895.004
##
## $tradeshare
##       Mean Standard Deviation  Minimum  Maximum
## 1 0.564703          0.2892703 0.140502 1.992616
```

```
##
## $yearsschool
##       Mean Standard Deviation Minimum Maximum
## 1 3.985077              2.542     0.2   10.07
##
## $rev_coups
##        Mean Standard Deviation Minimum   Maximum
## 1 0.1674501          0.2246798       0 0.9703704
##
## $assasinations
##        Mean Standard Deviation Minimum  Maximum
## 1 0.2775641          0.4915284       0 2.466667
```

b. Run a regression of $Growth$ on $TradeShare, YearsSchool, Rev_Coups, Assassinations$, and $RGDP60$. What is the value of the coefficient on Rev_Coups? Interpret the value of this coefficient. Is it large or small in a real-world sense?

Solution

```r
E62b <- lm(growth ~ tradeshare + yearsschool + rev_coups + assasinations + rgdp60, data = Growth)

summ(E62b)
```

```
## MODEL INFO:
## Observations: 65
## Dependent Variable: growth
## Type: OLS linear regression
##
## MODEL FIT:
## F(5,59) = 6.61, p = 0.00
## R² = 0.36
## Adj. R² = 0.30
##
## Standard errors: OLS
## ---------------------------------------------------
##                        Est.   S.E.   t val.      p
## ------------------- -------- ------ --------- ------
## (Intercept)            0.49   0.69     0.71   0.48
## tradeshare             1.56   0.76     2.06   0.04
## yearsschool            0.57   0.14     4.13   0.00
## rev_coups             -2.16   1.11    -1.94   0.06
## assasinations          0.35   0.48     0.74   0.46
## rgdp60                -0.00   0.00    -3.17   0.00
## ---------------------------------------------------
```

c. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

Solution

```r
# Sample means cross the sample linear regression.
mean(Growth$growth)
```

```
## [1] 1.942715
```

```r
# predict value
E62c <- function(x){
  E62b$coefficients %*% matrix(c(1, x), ncol = 1)
```

```r
}

E62c_X <- c(mean(Growth$tradeshare),
            mean(Growth$yearsschool),
            mean(Growth$rev_coups),
            mean(Growth$assasinations),
            mean(Growth$rgdp60))

E62c(E62c_X)
```

```
##          [,1]
## [1,] 1.942715
```

d. Repeat (c), but now assume that the country's value for *TradeShare* is one standard deviation above the mean.

Solution

```r
E62d_X <- c(mean(Growth$tradeshare) + sd(Growth$tradeshare),
            mean(Growth$yearsschool),
            mean(Growth$rev_coups),
            mean(Growth$assasinations),
            mean(Growth$rgdp60))

E62c(E62d_X)
```

```
##          [,1]
## [1,] 2.394468
```

e. Why is *Oil* omitted from the regression? What would happen if it were included?

Solution

The variable "oil" takes on the value of 0 for all 64 countries in the sample. This would generate perfect multicollinearity.

```r
sum(Growth$oil == 1)
```

```
## [1] 0
```