TA Session 6

Chi-Yuan Fang

April 13, 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	Omitted Variable, F test	04/14 Turn in HW3
7	04/06: No class			04/14 Turn in HW3
8	04/13: Class 6		Nonlinear	04/14 Turn in HW3, 04/20 Quiz 3
9	04/20: Class 7	Quiz 3	Review and Q&A	**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 8.1

> Lead is toxic, particularly for young children, and for this reason, government regulations severely restrict the amount of lead in our environment. But this was not always the case. In the early part of the 20th century, the underground water pipes in many U.S. cities contained lead, and lead from these pipes leached into drinking water. In this exercise, you will investigate the effect of these lead water pipes on infant mortality. On the text website http://www.pearsonglobaleditions.com, you will find the data file Lead_Mortality, which contains data on infant mortality, type of water pipes (lead or nonlead), water acidity (pH), and several demographic variables for 172 U.S. cities in 1900. A detailed description is given in **Lead_Mortality_Description**, also available on the website.
>
> a. Compute the average infant mortality rate (Inf) for cities with lead pipes and for cities with nonlead pipes. Is there a statistically significant difference in the averages?

Solution

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

Lead_Mortality <- read_excel("lead_mortality/lead_mortality.xlsx")

### average Inf for nonlead (pipe = 0) and lead pipes (pipe = 1)
tapply(Lead_Mortality$infrate, Lead_Mortality$lead, mean)
```

```
##         0         1 
## 0.3811679 0.4032576
```

The average infant mortality rate (Inf) for cities with lead pipes is 0.4033.

The average infant mortality rate (Inf) for cities with nonlead pipes is 0.3812.

Test for difference in mean:

- **Prepare**

  $H_0 : \overline{inf}_{lead} = \overline{inf}_{nonlead}$ v.s. $H_1 : \overline{inf}_{lead} \neq \overline{inf}_{nonlead}$

  Let the significance level be 5%.

- **Calculate**

```r
### data: lead pipes (pipe = 1)
Lead_Mortality_lead1 <- Lead_Mortality[Lead_Mortality$lead == 1,]

### data: nonlead (pipe = 0)
Lead_Mortality_lead0 <- Lead_Mortality[Lead_Mortality$lead == 0,]

### t test for difference
t.test(Lead_Mortality_lead1$infrate, Lead_Mortality_lead0$infrate,
       alternative = c("two.sided"),
       mu = 0, # H0
       var.equal = FALSE,
       conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  Lead_Mortality_lead1$infrate and Lead_Mortality_lead0$infrate
## t = 0.90387, df = 109.29, p-value = 0.3681
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.02634606  0.07052551
## sample estimates:
## mean of x mean of y
## 0.4032576 0.3811679
```

- **Conclusion**

  Because $p - value > \alpha = 5\%$, we do not reject H_0. There is no evidence that difference in mean is different from 0.

b. The amount of lead leached from lead pipes depends on the chemistry of the water running through the pipes. The more acidic the water is (that is, the lower its pH), the more lead is leached. Run a regression of Inf on $Lead, pH$, and the interaction term $Lead \times pH$.

   i. The regression includes four coefficients (the intercept and the three coefficients multiplying the regressors). Explain what each coefficient measures.
   ii. Plot the estimated regression function relating Inf to pH for $Lead = 0$ and for $Lead = 1$. Describe the differences in the regression functions, and relate these differences to the coefficients you discussed in (i).
   iii. Does $Lead$ have a statistically significant effect on infant mortality? Explain.
   iv. Does the effect of $Lead$ on infant mortality depend on pH? Is this dependence statistically significant?
   v. What is the average value of pH in the sample? At this pH level, what is the estimated effect of Lead on infant mortality? What is the standard deviation of pH? Suppose the pH level is one standard deviation lower than the average level of pH in the sample:

What is the estimated effect of *Lead* on infant mortality? What if pH is one standard deviation higher than the average value?

vi. Construct a 95% confidence interval for the effect of *Lead* on infant mortality when $pH = 6.5$.

Solution

i.

$$Inf = \beta_0 + \beta_1 lead + \beta_2 pH + \beta_3 lead \times pH + u \tag{1}$$

```
library(estimatr); library(car)

### regression
E81bi <- lm_robust(formula = infrate ~ lead + ph + lead*ph,
                   data = Lead_Mortality,
                   se_type = "stata")

summary(E81bi)
```

```
##
## Call:
## lm_robust(formula = infrate ~ lead + ph + lead * ph, data = Lead_Mortality,
##     se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)  0.91890    0.15049   6.106 6.866e-09  0.62180  1.21601 168
## lead         0.46180    0.20761   2.224 2.746e-02  0.05193  0.87167 168
## ph          -0.07518    0.02095  -3.588 4.369e-04 -0.11654 -0.03381 168
## lead:ph     -0.05686    0.02808  -2.025 4.448e-02 -0.11230 -0.00142 168
##
## Multiple R-squared:  0.2719 ,    Adjusted R-squared:  0.2589
## F-statistic: 20.97 on 3 and 168 DF,  p-value: 1.366e-11
```

- $\widehat{\beta}_0$, intercept, shows the level of Infrate when $lead = 0$ and $pH = 0$.
- $\widehat{\beta}_1$ and $\widehat{\beta}_3$ measure the effect of *lead* on the infant mortality rate. Other things being equal (pH is the same), the difference in predicted infant mortality rate between $lead = 1$ and $lead = 0$ is

$$\widehat{Inf}(lead = 1) - \widehat{Inf}(lead = 0) = 0.4618 - 0.0569pH. \tag{2}$$

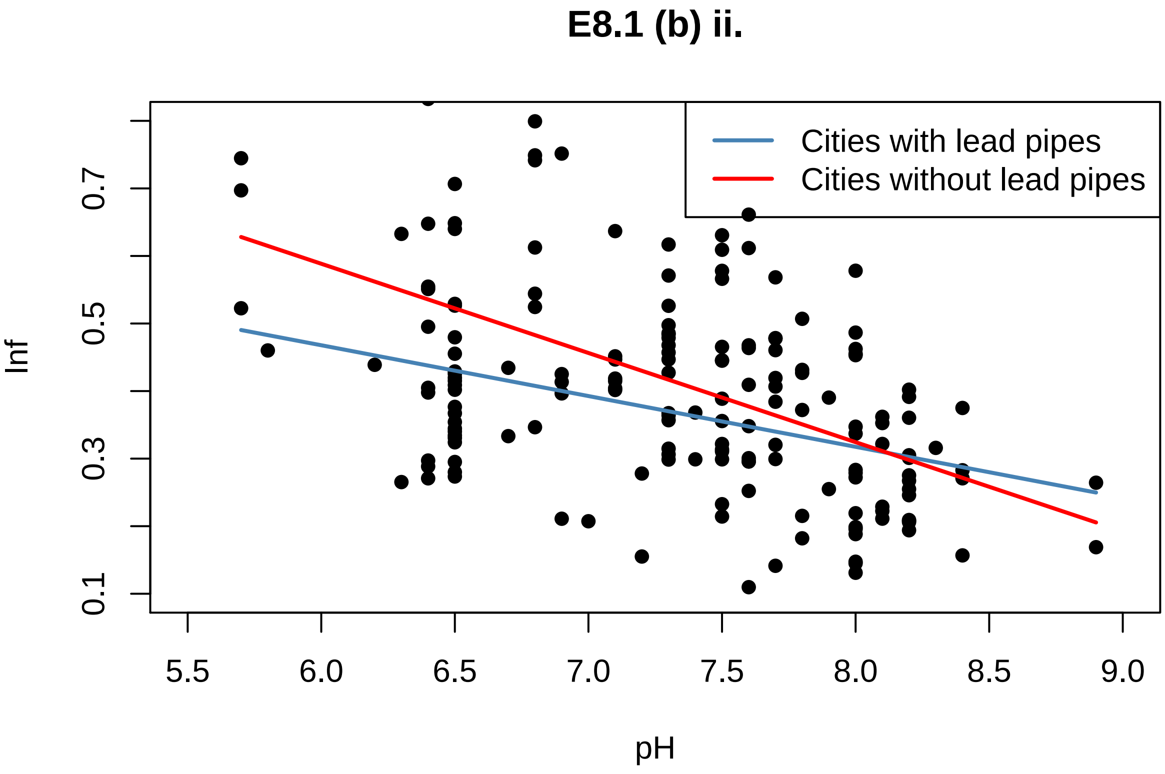
- $\widehat{\beta}_2$ and $\widehat{\beta}_3$ measure the effect of pH on the infant mortality rate. Other things being equal (*lead* is the same), the difference in predicted infant mortality rate between $pH = pH_0 + 1$ and $pH = pH_0$ is

$$\widehat{Inf}(pH = pH_0 + 1) - \widehat{Inf}(pH = pH_0) = -0.0752 - 0.0569lead. \tag{3}$$

ii.

```
### lead = 0
Yhat0 <- function(ph){
  E81bi$coefficients[1] + E81bi$coefficients[3]*ph
}
```

```r
### lead = 1
Yhat1 <- function(ph){
  E81bi$coefficients[1] + E81bi$coefficients[2] + E81bi$coefficients[3]*ph + E81bi$coefficients[4]*ph
}

plot(x = Lead_Mortality$ph,
     y = Lead_Mortality$infrate,
     pch = 16, # filled circle
     col = "black",
     xlim = c(5.5, 9),
     ylim = c(0.1, 0.8),
     xlab = "pH",
     ylab = "Inf",
     main = "E8.1 (b) ii.")

curve(Yhat0,
      from = min(Lead_Mortality$ph),
      to = max(Lead_Mortality$ph),
      add = TRUE,
      col = "steelblue",
      lwd = 2)

curve(Yhat1,
      from = min(Lead_Mortality$ph),
      to = max(Lead_Mortality$ph),
      add = TRUE,
      col = "red",
      lwd = 2)

# https://www.rdocumentation.org/packages/graphics/versions/3.6.2/topics/legend
legend("topright",
       legend = c("Cities with lead pipes", "Cities without lead pipes"),
       col = c("steelblue","red"),
       lty = 1,
       lwd = 2)
```



iii.

- **Prepare**

  $H_0 : \beta_{lead} = \beta_{lead \times pH} = 0$ v.s. H_1 : not H_0

  Let the significance level be 5%.

- **Calculate**

```
linearHypothesis(E81bi,
                 c("lead = 0", "lead:ph = 0"),
                 test = "F")
```

```
## Linear hypothesis test
##
## Hypothesis:
## lead = 0
## lead:ph = 0
##
## Model 1: restricted model
## Model 2: infrate ~ lead + ph + lead * ph
##
##   Res.Df Df     F  Pr(>F)
## 1    170
## 2    168  2 3.936 0.02135 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Conclusion**

  Because $p - value < \alpha = 5\%$, we reject H_0. There is significant evidence that *lead* has effect on Inf.

iv.

- **Prepare**

  $H_0 : \beta_{lead \times pH} = 0$ v.s. $H_1 : \beta_{lead \times pH} \neq 0$

  Let the significance level be 5%.

- **Calculate**

  See part (b) i.

- **Conclusion**

  Because $p - value < \alpha = 5\%$, we reject H_0. There is significant evidence that *lead* on Inf depend on pH.

v.

```
### average value of pH
pH_mu <- mean(Lead_Mortality$ph)
pH_mu
```

```
## [1] 7.322674
```

```
### estimated effect of Lead
E81bv <- function(pH){
  diff <- E81bi$coefficients[2] + E81bi$coefficients[4]*pH

  return(diff)
}

### estimated effect of Lead # average value of pH
E81bv(pH_mu)
```

```
##       lead
## 0.04541495
```

```
### standard deviation of pH
pH_sd <- sd(Lead_Mortality$ph)
pH_sd
```

```
## [1] 0.6917288
```

```
### estimated effect of Lead # pH_mu - pH_sd
E81bv(pH_mu - pH_sd)
```

```
##       lead
## 0.08474818
```

```
### estimated effect of Lead # pH_mu + pH_sd
E81bv(pH_mu + pH_sd)
```

```
##        lead
## 0.006081724
```

vi. Using Approach 2 of Section 7.3,

$$\begin{aligned}
Inf &= \beta_0 + \beta_1 lead + \beta_2 pH + \beta_3 lead \times pH + u \qquad (4)\\
&= \beta_0 + (\beta_1 lead + 0.65\beta_3 lead) + \beta_2 pH + (\beta_3 lead \times pH - 0.65\beta_3 lead) + u \qquad (5)\\
&= \beta_0 + b_1 lead + b_2 pH + b_3 lead \times (pH - 6.5) + u \qquad (6)
\end{aligned}$$

where

$$
\begin{aligned}
b_0 &= \beta_0 && (7)\\
b_1 &= \beta_1 + 0.65\beta_3 && (8)\\
b_2 &= \beta_2 && (9)\\
b_3 &= \beta_3. && (10)
\end{aligned}
$$

The effect of *Lead* on infant mortality is

$$\beta_1 + \beta_3 pH = b_1. \qquad (11)$$

```r
Lead_Mortality$phnew <- Lead_Mortality$ph - 6.5

E81bvi <- lm_robust(formula = infrate ~ lead + ph + lead*phnew,
                    data = Lead_Mortality,
                    se_type = "stata")

confint(E81bvi)[2,]
```

```
##      2.5 %     97.5 % 
## 0.02732276 0.15706533
```

95% confidence interval for the effect of *Lead* on infant mortality when $pH = 6.5$ is [0.0273, 0.1571].

c. The analysis in (b) may suffer from omitted variable bias because it neglects factors that affect infant mortality and that might potentially be correlated with *Lead* and *pH*. Investigate this concern, using the other variables in the data set.

Solution

We can check several demographic variables in the data set.

3 Empirical Exercise 8.2

On the text website http://www.pearsonglobaleditions.com, you will find a data file **CPS2015**, which contains data for full-time, full-year workers, ages 25–34, with a high school diploma or B.A./B.S. as their highest degree. A detailed description is given in **CPS2015_Description**, also available on the website. (These are the same data as in CPS96_15, used in Empirical Exercise 3.1, but are limited to the year 2015.) In this exercise, you will investigate the relationship between a worker's age and earnings. (Generally, older workers have more job experience, leading to higher productivity and higher earnings.)

a. ~ k.

Solution

Homework 4.

l. After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.

Solution

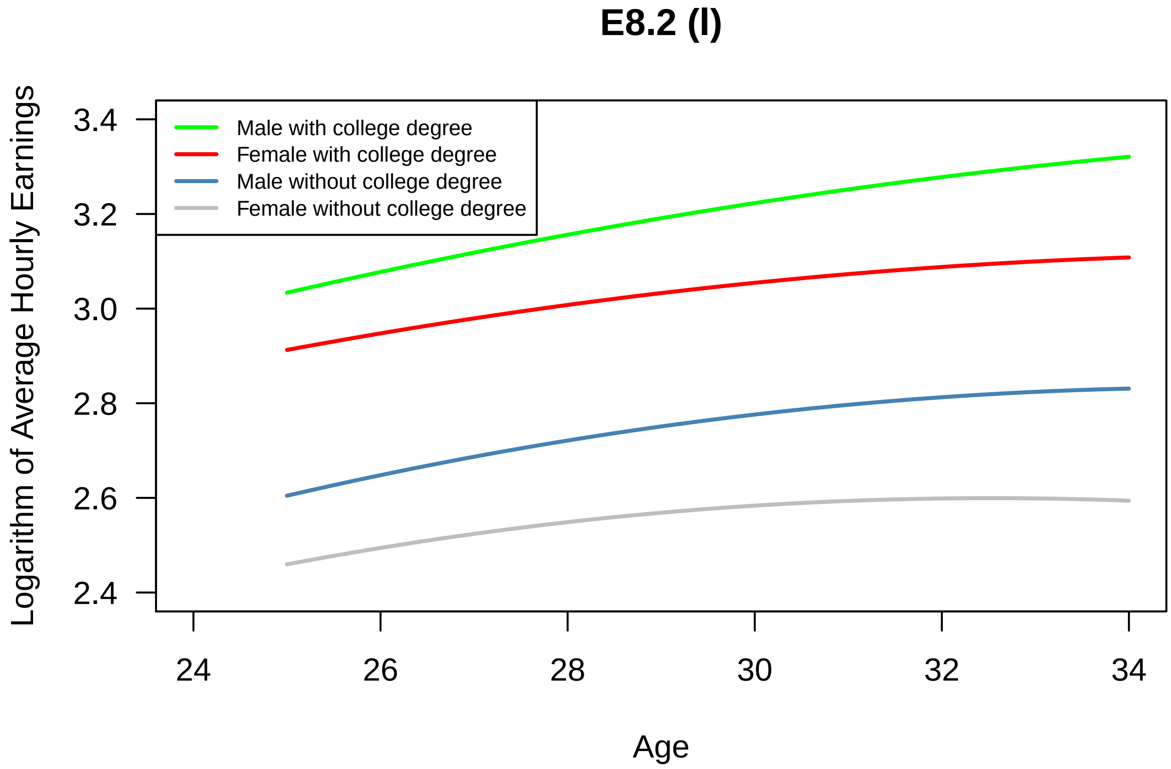
$$\ln AHE = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 female + \beta_4 bachelor + \beta_5 female \times bechelor + \beta_6 age * female + \beta_7 age^2 \times female + \beta_8 \qquad (12)$$

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

CPS2015 <- read_excel("CPS2015/CPS2015.xlsx")

### regression
E821 <- lm_robust(formula = log(ahe) ~ age + I(age^2) + female + bachelor + female*bachelor + age*femal
                  data = CPS2015,
                  se_type = "stata")

summary(E821)
```

```
##
## Call:
## lm_robust(formula = log(ahe) ~ age + I(age^2) + female + bachelor +
##     female * bachelor + age * female + I(age^2) * female + age *
##     bachelor + I(age^2) * bachelor, data = CPS2015, se_type = "stata")
##
## Standard error type:  HC1
##
## Coefficients:
##                     Estimate Std. Error   t value Pr(>|t|)  CI Lower  CI Upper
## (Intercept)        0.0302483   1.051127  0.028777  0.97704 -2.030274 2.091e+00
## age                0.1601922   0.071515  2.239971  0.02512  0.020001 3.004e-01
## I(age^2)          -0.0022888   0.001208 -1.894845  0.05815 -0.004657 7.906e-05
## female            -0.0399898   1.362979 -0.029340  0.97659 -2.711835 2.632e+00
## bachelor           0.9409250   1.362688  0.690492  0.48991 -1.730351 3.612e+00
## female:bachelor    0.0240159   0.022879  1.049706  0.29389 -0.020833 6.886e-02
## age:female         0.0002087   0.092842  0.002248  0.99821 -0.181789 1.822e-01
## I(age^2):female   -0.0001764   0.001570 -0.112383  0.91052 -0.003253 2.901e-03
## age:bachelor      -0.0405062   0.092752 -0.436714  0.66233 -0.222328 1.413e-01
## I(age^2):bachelor  0.0008013   0.001567  0.511249  0.60919 -0.002271 3.874e-03
##                     DF
## (Intercept)       7088
## age               7088
## I(age^2)          7088
## female            7088
## bachelor          7088
## female:bachelor   7088
## age:female        7088
## I(age^2):female   7088
## age:bachelor      7088
## I(age^2):bachelor 7088
##
## Multiple R-squared:  0.2101 ,    Adjusted R-squared:  0.2091
## F-statistic: 212.9 on 9 and 7088 DF,  p-value: < 2.2e-16
```

```r
plot(x = CPS2015$age,
     y = log(CPS2015$ahe),
     pch = 16, # filled circle
     col = "white",
     xlim = c(24, 34),
     ylim = c(2.4, 3.4),
     xlab = "Age",
     ylab = "Logarithm of Average Hourly Earnings",
     main = "E8.2 (l)",
     las = 1)

### male (female = 0) with college degree (bachelor = 1)
Yhat1 <- function(age){
  Fe <- 0; BA <- 1

  E821$coefficients[1] + E821$coefficients[2]*age + E821$coefficients[3]*age^2 + E821$coefficients[4]*F
}

curve(Yhat1,
      from = min(CPS2015$age),
      to = max(CPS2015$age),
      add = TRUE,
      col = "green",
      lwd = 2)

### female (female = 1) with college degree (bachelor = 1)
Yhat2 <- function(age){
  Fe <- 1; BA <- 1

  E821$coefficients[1] + E821$coefficients[2]*age + E821$coefficients[3]*age^2 + E821$coefficients[4]*F
}

curve(Yhat2,
      from = min(CPS2015$age),
      to = max(CPS2015$age),
      add = TRUE,
      col = "red",
      lwd = 2)

### male (female = 0) without college degree (bachelor = 0)
Yhat3 <- function(age){
  Fe <- 0; BA <- 0

  E821$coefficients[1] + E821$coefficients[2]*age + E821$coefficients[3]*age^2 + E821$coefficients[4]*F
}

curve(Yhat3,
      from = min(CPS2015$age),
      to = max(CPS2015$age),
      add = TRUE,
      col = "steelblue",
      lwd = 2)

### female (female = 1) without college degree (bachelor = 0)
Yhat4 <- function(age){
  Fe <- 1; BA <- 0

  E821$coefficients[1] + E821$coefficients[2]*age + E821$coefficients[3]*age^2 + E821$coefficients[4]*F
}
```

```
curve(Yhat4,
      from = min(CPS2015$age),
      to = max(CPS2015$age),
      add = TRUE,
      col = "gray",
      lwd = 2)

legend("topleft",
       legend = c("Male with college degree", "Female with college degree", "Male without college degre
       col = c("green", "red", "steelblue", "gray"),
       lty = 1,
       lwd = 2,
       cex = 0.7)
```


E8.2 (I)
Male with college degree
Female with college degree
Male without college degree
Female without college degree
Logarithm of Average Hourly Earnings
Age