Empirical Exercise - E8.1

Chi-Yuan Fang

2021-04-03

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

```
### 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**

```
### 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.

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

```
### regression
E81bi <- lm(formula = infrate ~ lead + ph + lead*ph,
            data = Lead_Mortality)

### output table
library(stargazer)

stargazer(E81bi,
          type = "text",
          digits = 4)
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                infrate
## ------------------------------------------------
## lead                          0.4618**
##                               (0.2212)
##
## ph                           -0.0752***
##                               (0.0243)
##
## lead:ph                       -0.0569*
##                               (0.0304)
##
## Constant                      0.9189***
##                               (0.1745)
##
## ------------------------------------------------
## Observations                     172
## R2                              0.2719
## Adjusted R2                     0.2589
## Residual Std. Error      0.1303 (df = 168)
## F Statistic          20.9083*** (df = 3; 168)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

- $\widehat{\beta}_0$, 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

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

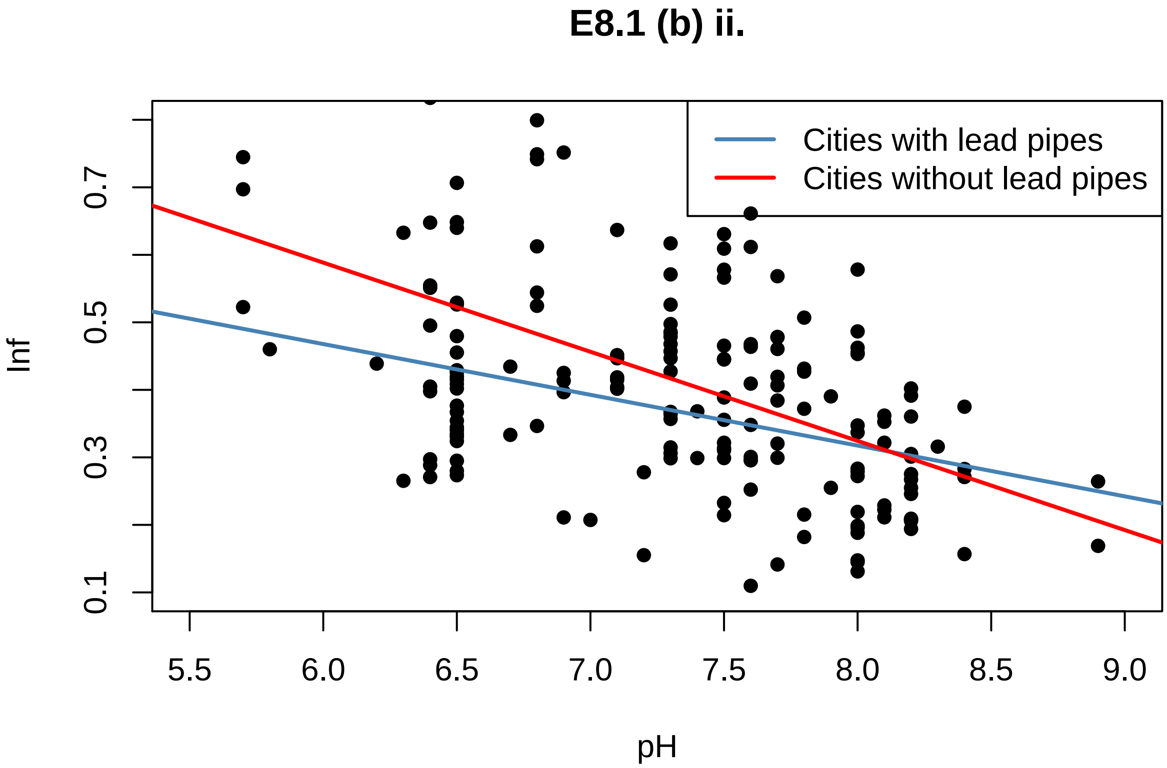
- $\widehat{\beta}_2$ and $\widehat{\beta}_3$ measure the effect of pH on the infant mortality rate. Other things being equal (*lead* is the same), the difference in predicted infant mortality rate between $pH = pH_0 + 1$ and $pH = pH_0$ is

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

ii.

```
### lead = 0
E81bii0 <- lm(formula = infrate ~ ph,
              data = Lead_Mortality_lead0)

### lead = 1
E81bii1 <- lm(formula = infrate ~ ph,
              data = Lead_Mortality_lead1)

plot(x = Lead_Mortality$ph,
     y = Lead_Mortality$infrate,
     pch = 16, # filled circle
     col = "black",
```

```r
     xlim = c(5.5, 9),
     ylim = c(0.1, 0.8),
     xlab = "pH",
     ylab = "Inf",
     main = "E8.1 (b) ii.")

abline(E81bii0, lwd = 2, col = "steelblue")

abline(E81bii1, lwd = 2, col = "red")

# 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**

```r
library(car)

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    RSS Df Sum of Sq      F  Pr(>F)
## 1    170 3.0012
## 2    168 2.8512  2      0.15 4.4191 0.01348 *
## ---
## 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.

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

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

```r
### 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
```

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

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

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

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

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

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

vi. Using Approach 2 of Section 7.3,

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

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

$$\beta_1 + \beta_3 pH = b_1. \tag{11}$$

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

E81bvi <- lm(formula = infrate ~ lead + ph + lead*phnew,
             data = Lead_Mortality)

confint(E81bvi)[2,]
```

```
##     2.5 %    97.5 %
## 0.0304741 0.1539140
```

95% confidence interval for the effect of *Lead* on infant mortality when $pH = 6.5$ is [0.0305, 0.1539].

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 dataset.