Empirical Exercise - E5.1

Chi-Yuan Fang

2021-03-21

Use the data set **Earnings_and_Height** described in Empirical Exercise 4.2 to carry out the following exercises.

a. Run a regression of $Earnings$ on $Height$.
   i. Is the estimated slope statistically significant?
   ii. Construct a 95% confidence interval for the slope coefficient.

Solution

i.

- **Prepare**

$H_0 : \beta_{height} = 0$ v.s. $H_1 : \beta_{height} \neq 0$

Let the significance level be 0.05.

- **Calculate**

```r
# import data
library(readxl)
Earnings_and_Height <- read_xlsx("Earnings_and_Height/Earnings_and_Height.xlsx")

E51a_model <- lm(earnings ~ height, data = Earnings_and_Height)

summary(E51a_model)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = Earnings_and_Height)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -512.73    3386.86  -0.151     0.88
## height        707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

- **Conclude**

Because $p - value < 0.05$, we reject H_0. The estimated slope is statistically significant different from 0.

ii.

In simple regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{1}$$

we know

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \tag{2}$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \tag{3}$$

$$ESS = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2 \tag{4}$$

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \tag{5}$$

$$SSR = \sum_{i=1}^{n} \widehat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2 \tag{6}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{7}$$

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} \widehat{u}_i^2}{n-2} \tag{8}$$

$$SE\left(\widehat{\beta}_1\right) = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}} \tag{9}$$

$$SE\left(\widehat{\beta}_0\right) = \sqrt{\widehat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}\right)} \tag{10}$$

```r
E51a <- function(x, y){
  # numbers of sample
  n <- length(y)

  # sample mean
  xbar <- mean(x)
  ybar <- mean(y)

  # OLS coefficient
  b1hat <- cov(x,y)/var(x)
  b0hat <- ybar - b1hat*xbar

  yhat <- b0hat + b1hat*x

  # explained sum of squares (ESS)
  ESS <- sum((yhat - ybar)^2)
  # total sum of squares (TSS)
  TSS <- sum((y - ybar)^2)
  # sum of squared residuals (SSR)
  SSR <- sum((y - yhat)^2)

  # coefficient of determination
```

```r
  Rsqure <- ESS/TSS

  # standard error of the regression
  SER <- sqrt(SSR/(n-2))

  # standard error of coefficient
  se_b1hat <- sqrt(SER^2/sum((x - xbar)^2))
  se_b0hat <- sqrt(SER^2* (1/n + xbar^2/sum((x - xbar)^2)))

  # 95% CI for b1hat
  lower_b1hat <- round(b1hat - qnorm(0.975, mean = 0, sd = 1)*se_b1hat, digit = 4)
  upper_b1hat <- round(b1hat + qnorm(0.975, mean = 0, sd = 1)*se_b1hat, digit = 4)
  CI_b1hat <- paste(lower_b1hat, "-"  ,upper_b1hat)

  # 95% CI for b0hat
  lower_b0hat <- round(b0hat - qnorm(0.975, mean = 0, sd = 1)*se_b0hat, digit = 4)
  upper_b0hat <- round(b0hat + qnorm(0.975, mean = 0, sd = 1)*se_b0hat, digit = 4)
  CI_b0hat <- paste(lower_b0hat, "-"  ,upper_b0hat)

  # coefficient
  coef <- matrix(c(b0hat, se_b0hat, CI_b0hat, b1hat, se_b1hat, CI_b1hat), ncol = 3, byrow = TRUE)
  rownames(coef) <- c("Intercept", "Slope")
  colnames(coef) <- c("Estimate", "Standard Error", "95% Confidence Interval")

  result <- list(coef, Rsqure)
  names(result) <- c("Coefficients", "R-squared")

  result
}

E51a(Earnings_and_Height$height, Earnings_and_Height$earnings)
```

```
## $Coefficients
##           Estimate             Standard Error     95% Confidence Interval
## Intercept "-512.733592001881" "3386.85615092263" "-7150.8497 - 6125.3825"
## Slope     "707.671558437266"  "50.4892245961979" "608.7145 - 806.6286"
##
## $`R-squared`
## [1] 0.0108753
```

b. Repeat (a) for women.

Solution

i.

- **Prepare**

$H_0 : \beta^{women}_{height} = 0$ v.s. $H_1 : \beta^{women}_{height} \neq 0$

Let the significance level be 0.05.

- **Calculate**

```r
# data: woman
Earnings_and_Height_women <- Earnings_and_Height[Earnings_and_Height$sex == 0,]

E51b_model <- lm(earnings ~ height, data = Earnings_and_Height_women)
```

```r
summary(E51b_model)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = Earnings_and_Height_women)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42748 -22006  -7466  36641  46865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12650.9     6383.7   1.982   0.0475 *
## height         511.2       98.9   5.169   2.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26800 on 9972 degrees of freedom
## Multiple R-squared:  0.002672,   Adjusted R-squared:  0.002572
## F-statistic: 26.72 on 1 and 9972 DF,  p-value: 2.396e-07
```

- **Conclude**

Because $p-value < 0.05$, we reject H_0. The estimated slope is statistically significant different from 0.

ii.

```r
E51a(Earnings_and_Height_women$height, Earnings_and_Height_women$earnings)
```

```
## $Coefficients
##           Estimate           Standard Error     95% Confidence Interval
## Intercept "12650.8577295031" "6383.74100734725" "138.9553 - 25162.7602"
## Slope     "511.222170015359" "98.8963075918224" "317.389 - 705.0554"
##
## $`R-squared`
## [1] 0.002672482
```

c. Repeat (a) for men.

Solution

i.

- **Prepare**

$H_0 : \beta^{men}_{height} = 0$ v.s. $H_1 : \beta^{women}_{height} \neq 0$

Let the significance level be 0.05.

- **Calculate**

```r
# data: woman
Earnings_and_Height_men <- Earnings_and_Height[Earnings_and_Height$sex == 1,]

E51c_model <- lm(earnings ~ height, data = Earnings_and_Height_men)

summary(E51c_model)
```

```
##
## Call:
```

```
## lm(formula = earnings ~ height, data = Earnings_and_Height_men)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -50158 -22373  -8118  33091  59228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43130.3     7068.5  -6.102  1.1e-09 ***
## height        1306.9      100.8  12.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26670 on 7894 degrees of freedom
## Multiple R-squared:  0.02086,    Adjusted R-squared:  0.02074
## F-statistic: 168.2 on 1 and 7894 DF,  p-value: < 2.2e-16
```

- **Conclude**

Because $p-value < 0.05$, we reject H_0. The estimated slope is statistically significant different from 0.

ii.

```
E51a(Earnings_and_Height_men$height, Earnings_and_Height_men$earnings)
```

```
## $Coefficients
##           Estimate             Standard Error     95% Confidence Interval
## Intercept "-43130.3423470527" "7068.48053034493" "-56984.3096 - -29276.3751"
## Slope     "1306.85990584335"  "100.766159761487" "1109.3619 - 1504.3579"
##
## $`R-squared`
## [1] 0.02086292
```

d. Test the null hypothesis that the effect of height on earnings is the same for men and women. (Hint: See Exercise 5.15.)

Solution

- **Prepare**

$H_0 : \beta^{men}_{height} - \beta^{women}_{height} = 0$ v.s. $H_1 : \beta^{men}_{height} - \beta^{women}_{height} \neq 0$

Let the significance level be 0.05.

- **Calculate**

```
E51d <- function(x1, y1, x2, y2){
  # numbers of sample
  n1 <- length(y1); n2 <- length(y2)

  # sample mean
  x1bar <- mean(x1); x2bar <- mean(x2)
  y1bar <- mean(y1); y2bar <- mean(y2)

  # OLS coefficient
  b1hat1 <- cov(x1, y1)/var(x1); b1hat2 <- cov(x2, y2)/var(x2)
  b0hat1 <- y1bar - b1hat1*x1bar; b0hat2 <- y2bar - b1hat2*x2bar

  y1hat1 <- b0hat1 + b1hat1*x1; y2hat2 <- b0hat2 + b1hat2*x2
```

```
  # explained sum of squares (ESS)
  ESS1 <- sum((y1hat1 - y1bar)^2); ESS2 <- sum((y2hat2 - y2bar)^2)
  # total sum of squares (TSS)
  TSS1 <- sum((y1 - y1bar)^2); TSS2 <- sum((y2 - y2bar)^2)
  # sum of squared residuals (SSR)
  SSR1 <- sum((y1 - y1hat1)^2); SSR2 <- sum((y2 - y2hat2)^2)

  # standard error of the regression
  SER1 <- sqrt(SSR1/(n1-2)); SER2 <- sqrt(SSR2/(n2-2))

  # standard error of coefficient
  se_b1hat1 <- sqrt(SER1^2/sum((x1 - x1bar)^2))
  se_b1hat2 <- sqrt(SER2^2/sum((x2 - x2bar)^2))

  est <- b1hat1 - b1hat2
  se <- sqrt(se_b1hat1^2 + se_b1hat2^2)

  # 95% CI for difference
  lower <- round(est - qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  upper <- round(est + qnorm(0.975, mean = 0, sd = 1)*se, digit = 4)
  CI <- paste(lower, "-", upper)

  # output table
  Table <- data.frame(est, se, CI)
  colnames(Table) <- c("Estimate", "Standard Error", "95% Confidence Interval")

  Table

}

E51d(Earnings_and_Height_men$height, Earnings_and_Height_men$earnings, Earnings_and_Height_women$height
```

```
##   Estimate Standard Error 95% Confidence Interval
## 1 795.6377       141.1889    518.9126 - 1072.3628
```

- **Conclude**

Because $0 \notin 95\%$ confidence interval, we reject H_0. The estimated slope is statistically significant different from 0.

e. One explanation for the effect of height on earnings is that some professions require strength, which is correlated with height. Does the effect of height on earnings disappear when the sample is restricted to occupations in which strength is unlikely to be important?

Solution

```
E51e <- function(x, y){
  # numbers of sample
  n <- length(y)

  # sample mean
  xbar <- mean(x)
  ybar <- mean(y)

  # OLS coefficient
  b1hat <- cov(x,y)/var(x)
```

```r
  b0hat <- ybar - b1hat*xbar

  yhat <- b0hat + b1hat*x

  # explained sum of squares (ESS)
  ESS <- sum((yhat - ybar)^2)
  # total sum of squares (TSS)
  TSS <- sum((y - ybar)^2)
  # sum of squared residuals (SSR)
  SSR <- sum((y - yhat)^2)

  # coefficient of determination
  #Rsqure <- ESS/TSS

  # standard error of the regression
  SER <- sqrt(SSR/(n-2))

  # standard error of coefficient
  se_b1hat <- sqrt(SER^2/sum((x - xbar)^2))

  # 95% CI for b1hat
  lower_b1hat <- round(b1hat - qnorm(0.975, mean = 0, sd = 1)*se_b1hat, digit = 4)
  upper_b1hat <- round(b1hat + qnorm(0.975, mean = 0, sd = 1)*se_b1hat, digit = 4)
  CI_b1hat <- paste(lower_b1hat, "-"  ,upper_b1hat)

  Table <- matrix(c(b1hat, se_b1hat, CI_b1hat), nrow = 1)
  colnames(Table) <- c("Estimate", "Standard Error", "95% Confidence Interval")

  Table
}

E51e_output <- matrix(nrow = 15, ncol = 3)

rownames(E51e_output) <- c(1:15)
colnames(E51e_output) <- c("Estimate", "Standard Error", "95% Confidence Interval")

for (i in 1:15){
  data <- Earnings_and_Height[Earnings_and_Height$occupation == i,]
  x <- data$height
  y <- data$earnings
  E51e_output[i,] <- E51e(x, y)
}

E51e_output
```

```
##    Estimate             Standard Error     95% Confidence Interval
## 1  "469.458070315024"  "155.197956234637" "165.2757 - 773.6405"
## 2  "622.755176016371"  "117.270342579588" "392.9095 - 852.6008"
## 3  "649.721773596878"  "214.630174152533" "229.0544 - 1070.3892"
## 4  "1372.384840781"    "148.798366501132" "1080.7454 - 1664.0243"
## 5  "201.215856644732"  "133.739773995006" "-60.9093 - 463.341"
## 6  "-172.893730327361" "680.403429299889" "-1506.4599 - 1160.6725"
## 7  "1503.03853605894"  "403.078327473455" "713.0195 - 2293.0575"
## 8  "62.8574693717754"  "128.266422708668" "-188.5401 - 314.255"
```

```
## 9  "1049.20129254131"  "308.815799047297" "443.9334 - 1654.4691"
## 10 "571.223149958568"  "331.525682080739" "-78.5552 - 1221.0015"
## 11 "967.009049351859"  "306.481847438174" "366.3157 - 1567.7024"
## 12 "1080.32127574574"  "286.488509387146" "518.8141 - 1641.8284"
## 13 "972.909550457317"  "150.635402614451" "677.6696 - 1268.1495"
## 14 "1138.41360943808"  "268.02788099113"  "613.0886 - 1663.7386"
## 15 "549.115247105915"  "249.240665896231" "60.6125 - 1037.618"
```