Econometric Methods

Solutions to empirical exercise 5.1, Chapter 5 S&W

Zaeen de Souza

14/07/2021

Contents

Background: Empirical Exercise 5.1

These are the solutions to **E5.1** from **Chapter 5** of *Introduction to Econometrics (Updated Third edition)* by Stock & Watson. You should have the following on your computer in order to check answers/run the code and follow the questions in this assignment:

- An updated version of `R` and `Rstudio`.
- The following packages installed:
  - `ggplot2`
  - `readxl`
  - `stargazer`
- The dataset called `height_and_earnings`.
- The data description pdf to understand the variables being used.

Reading guide

All the code needed to complete the assignments is within this document. R code will be in a grey box and will look like this:

```
summary(iris)
```

And all R output i.e what R shows you once you run some code, will have `#` signs next to it, and will look like this:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.300   5.100   5.800   5.843   6.400   7.900
```

As far as possible these guides will show the **exact output** that comes from running code in `R`, and at times will use formatted tables made in latex. The results themselves, will be identical. Some things to note, that might make output look different accross different computers:

- R reports things like p-values using scientific notation, but some computers report the numbers with many trailing zeros.

- If you have an old version of `R` or `Rstudio` it is highly recommended that you update it using the following code:

```
# Use this to update R from within RStudio
install.packages("installr")
library(installr)
# This last command, will open up a download prompt; choose yes/no accordingly.
updateR()
```

For updating `Rstudio`, un-install your version of `RStudio`, and download a fresh version from the `RStudio` website.

Loading the data and libraries

The following code sets the working directory to the folder where you have downloaded the data, loads the libraries needed fo the assignment and loads the excel dataset.

```
# Loading excel files
library(readxl)
# Making graphs
library(ggplot2)

# Setting working directory - this is unique to your computer
setwd("~Zaeen de Souza/Chapter 5 Hypo. Testing and CI Single Reg")

# Loading the data as 'pset_data'
pset_data <- read_excel("Earnings_and_Height.xlsx")
```

Exercise E5.1

A. Run a regression of Earnings on Height.

```
model_1 <- lm(earnings ~ height, data = pset_data)
summary(model_1)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = pset_data)
##
## 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
```

A.i Is the estimated slope statistically significant?

The slope is statistically significant. Under the null hypothesis of no effect (i.e $\beta_{H_0} = 0$), we can calculate the T-statistic for the estimated slope using:

$$T_{actual} = \frac{\hat{\beta}_1 - \beta_{H_0}}{SE(\hat{\beta})}, \quad where,\ \beta_{H_0} = 0$$

This gives a T-statistic of 14.016 and the p-value is < 0.01, which is enough information for us to reject H_0, i.e the null that the estimated slope $\hat{\beta}_1 = \beta_{H_0} = 0$, and accept H_1 i.e the alternative hypothesis that $\hat{\beta} \neq 0$.

You can manually calculate the t-statistic by

```
# retrieving the stored coefficient from the saved model object
# dividing it by the saved standard error -
# Note: it retains the name of the variable (height)
t <- round(model_1$coefficients[2]/coef(summary(model_1))[, "Std. Error"][2],3)
t
```

```
## height
## 14.016
```

A.ii Construct a 95% confidence interval for the slope coefficient.

Given the estimated slope which is 707.67 and the standard error of 50.49, we can construct the 95% confidence interval using the following formula

$$\hat{\beta} \pm (1.96 \times SE(\hat{\beta}))$$

Where $\hat{\beta}$ is the estimated slope, 1.96 is the critical value from the standard normal distribution and $SE(\hat{\beta})$ is the standard error of the estimated slope. Taking the values directly from the regression output, we have,

$$707.67 \pm (1.96 \times 50.49) = [608.71, 806.63]$$

The R code to automate the calculation of the interval is below. You can verify that the hand calculation matches the R output by,

```
confint(model_1)
```

```
##                  2.5 %    97.5 %
## (Intercept) -7151.2994 6125.8322
## height        608.7078  806.6353
```

B. Repeat (a) for women.

```
model_2 <- lm(earnings ~ height, data = subset(pset_data, sex == 0))
summary(model_2)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = subset(pset_data, sex ==
##     0))
##
## 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
```

B.i Is the estimated slope statistically significant?

The slope is statistically significant—looking at p-value for the t-test which is < 0.01, we can reject the null that the estimated slope for female workers is equal to 0.

B. ii Construct a 95% confidence interval for the slope coefficient.

Using the formula from `A.ii`, we have the following 95% confidence interval for our estimated slope for female workers.

$$511.2 \pm (1.96 \times 98.9) = [317.36, 705.04]$$

```
confint(model_2)
```

```
##                 2.5 %     97.5 %
## (Intercept) 137.4364 25164.2790
## height      317.3654   705.0789
```

C. Repeat (a) for men.

```
model_3 <- lm(earnings ~ height, data = subset(pset_data, sex == 1))
summary(model_3)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = subset(pset_data, sex ==
##     1))
##
## 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
```

C.i Is the estimated slope statistically significant?

The slope is statistically significant—looking at p-value for the t-test which is < 0.01, we can reject the null hypothesis that the estimated slope for male workers is equal to 0.

C. ii Construct a 95% confidence interval for the slope coefficient.

Using the formula from `A.ii`, we have the following 95% confidence interval for our estimated slope for male workers,

$$1306.9 \pm (1.96 \times 100.8) = [1109.33, 1504.47]$$

```
confint(model_3)
```

```
##                  2.5 %     97.5 %
## (Intercept) -56986.434 -29274.251
## height        1109.332   1504.388
```

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

Let the difference between the two coefficients be given by,

$$\hat{\beta}_{\Delta} = \hat{\beta}_m - \hat{\beta}_f$$

Where the subscripts m and f denote male and female, respectively. Recall that the standard error of $\hat{\beta}_{\Delta}$ is given by the following formula.:

$$SE(\hat{\beta}_{\Delta}) = \sqrt{[SE(\hat{\beta}_m)]^2) + [SE(\hat{\beta}_f)]^2}$$

Note as well, some people might find this version easier to work with, which has been slightly re-arranged—both will give you the same answer

$$SE(\hat{\beta}_{\Delta}) = ([SE(\hat{\beta}_m)]^2) + [SE(\hat{\beta}_f)]^2)^{\frac{1}{2}}$$

Substituting, values from the previous questions, we have the difference in coefficients given by,

$$\hat{\beta}_{\Delta} = (1306.9 - 511.2) = 795.7$$

And using the formula to calculate the standard error of the difference, again, using the estimated coefficients, we have the following:

$$SE(\hat{\beta}_{\Delta}) = \sqrt{[(100.8^2) + (98.9^2)]} = 101.2$$

Therefore it follows from the standard formula for the confidence interval, that the 95% confidence interval within which the difference falls is given by,

$$795.7 \pm (1.96 \times 101.2) = [597.3, 994.1]$$

E. One explanation for the effect on 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?

We use the subset function again, to run our regression on a subset of the data for occupations that are likely to not have a major need for strength. Note the use of the `'|'` operator that allows us to use an `'or'` condition within the subset function[1].

```
model_4 <- lm(earnings ~ height,
              data = subset(pset_data,
                            occupation == 1 |
                            occupation == 2 |
                            occupation == 4 |
                            occupation == 5 |
                            occupation == 6 |
                            occupation == 8))
summary(model_4)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = subset(pset_data, occupation ==
##     1 | occupation == 2 | occupation == 4 | occupation == 5 |
##     occupation == 6 | occupation == 8))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52992 -22858  -8120  32158  52535
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15056.21    4259.18  -3.535 0.000409 ***
## height         970.33      64.06  15.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27230 on 12236 degrees of freedom
## Multiple R-squared:  0.01841,    Adjusted R-squared:  0.01833
## F-statistic: 229.5 on 1 and 12236 DF,  p-value: < 2.2e-16
```

The correlation between height and wages does not disappear when restricting the analysis only to people who work in the following occupations.

- 1 = Exec/Manager
- 2 = Professionals
- 4 = Sales
- 5 = Administration

[1] Note as well, that this is just one method of using multiple conditions in the subset function. You could also experiment with functions such as %in%

- 6 = Household service
- 8 = Other Service

In fact, the estimated slope is larger than the coefficient from the full sample regression.