Econometric Methods

Solutions to empirical exercise 4.2, Chapter 4 S&W

Zaeen de Souza

13/07/2021

Contents

Background: Empirical Exercise 4.2

These are the solutions to **E4.2** from **Chapter 4** 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 for the assignment and loads the excel dataset.

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

# Setting working directory - this is unique to your computer

setwd("~Zaeen de Souza/Datasets/Chapter 4 Single Regression")

# Loading the data as 'pset_data'

pset_data <- read_excel("Earnings_and_Height.xlsx")
```

Exercise E4.2

A. What is the median value of height in the sample?

The median height of the sample is 67 inches.

```
summary(pset_data$height)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   48.00   64.00   67.00   66.96   70.00   84.00
```

B

B.i. Estimate average earnings for workers whose height is at most 67 inches.

The mean earnings for a worker with height ≤ 67 is $44488.44. `na.rm=TRUE` tells R to omit missing values from the function—without this, it will evaluate to NA if there are missing values.

```
below <- subset(pset_data, height <= 67)
mean(below$earnings, na.rm = T)
```

```
## [1] 44488.44
```

B.ii. Estimate average earnings for workers whose height is greater than 67 inches.

The mean earnings for a worker with height > 67 is $49987.88.

```
above <- subset(pset_data, height > 67)
mean(above$earnings, na.rm = T)
```

```
## [1] 49987.88
```

B.iii. On average, do taller workers earn more than shorter workers? How much more? What is a 95% confidence interval for the difference in average earnings?

We answer this with a T-test, using the data subsets we saved in `B.i` and `B.ii`.

```
t.test(above$earnings, below$earnings)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  above$earnings and below$earnings
## t = 13.59, df = 16624, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  4706.237 6292.643
## sample estimates:
## mean of x mean of y 
##  49987.88  44488.44
```

The difference in means is $49987.88 - 44488.44 = 5499.44$. From the output we see that the p-value is smaller than 0.01, so we reject the null that true difference is 0 in favour of the alternative. The output also shows us that the estimated difference has a 95% confidence interval of $[4706.237, 6292.643]$.

If you want to manually see what is happening inside the `t.test` function, you can replicate the observed T-statistic using the following formula:

$$T_{actual} = \frac{(\hat{Y}_{above} - \hat{Y}_{below}) - d_0}{\sqrt{\frac{\hat{\sigma}^2_{above}}{n_{above}} + \frac{\hat{\sigma}^2_{below}}{n_{below}}}}$$

Where the $\hat{Y}$ is the estimated mean earnings, groups are specified using the two subscripts. The denominator is the standard error of the difference in the two means and d_0 is the difference as hypothesised under the null, which is 0 in this case.

```r
# d0 equal to zero since we don't want to impose a direction on the test
d0 <- 0

# here we create the mean, variance and n of the above sample
a_mu <- mean(above$earnings, na.rm = T)
a_var <- var(above$earnings, na.rm = T)
a_n <- length(above$earnings)

# here we create the mean, variance and n of the below sample
b_mu <- mean(below$earnings, na.rm = T)
b_var <- var(below$earnings, na.rm = T)
b_n <- length(below$earnings)

# rounding up to 2 decimal places with round(formula,2)
round((a_mu-b_mu-d0)/(sqrt((a_var/a_n-1)+(b_var/b_n-1))), 2)
```

```
## [1] 13.59
```

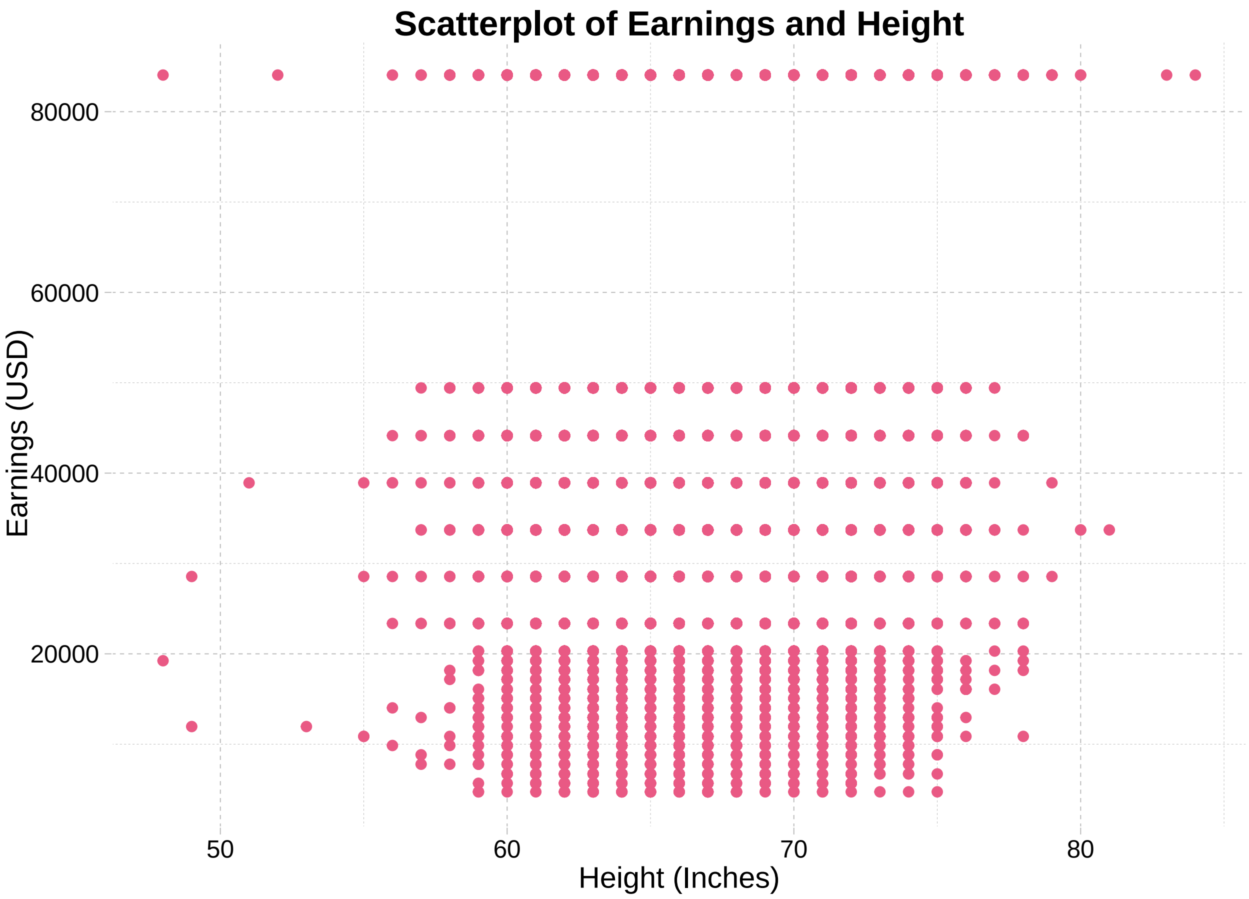
You can cross-check the T-statistic from the hand calculation and from the `R` function. They are the same.

C.

C.i. Construct a scatterplot of annual earnings (Earnings) on height(Height). Notice that the points on the plot fall along horizontal lines. (There are only 23 distinct values of Earnings). Why? (Hint: Carefully read the detailed data description.)

These are estimated mean earnings that are grouped into 23 income brackets and so we see 23 horizontal lines in the plot. The following code produces a scatterplot[1]. Note clustering of points along horizontal lines, especially in the <20000 range.

```
ggplot(pset_data) +
    geom_point(aes(y = earnings,
                   x = height),
               colour = "#e95984") +
    ylab("Earnings (USD)") +
    xlab("Height (Inches)") +
  labs(title = "Scatterplot of Earnings and Height")
```



[1] An aside: you can experiment with the 'ggplot' function to change the look of the plot

D. Run a regression of Earnings on Height.

We need to estimate the following linear model using OLS.

$$Earnings_i = \alpha + \beta_1 Height_i + \varepsilon_i$$

The slope is given by the estimate of β_1 i.e $\hat{\beta_1}$. The following code estimates the regression.

```r
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
```

D.i. What is the estimated slope?

The slope i.e the $\hat{\beta_1}$ is 707.67. A one inch increase in height is associated with, on average, a 707.67 dollar increase in annual earnings.

D.ii. Use the estimated regression to predict earnings for a worker who is 67 inches tall, for a worker who is 70 inches tall, and for a worker who is 65 inches tall.

For a worker who is 67 inches tall, the equation becomes,

$$-512.73 + (707.67 \times 67) = 46901.16$$

For a worker who is 70 inches tall, we have,

$$-512.73 + (707.67 \times 70) = 49024.17$$

And for a worker who is 65 inches tall, we have,

$$-512.73 + (707.67 \times 65) = 45485.82$$

E. Suppose height were measured in centimeters instead of inches. Answer the following questions about the Earnings on Height (in cm) regression.

We will create a new variable called `cm_height`, and then re-run the regression we ran in question D.

```
pset_data$cm_height <- pset_data$height * 2.54
model_2 <- lm(earnings ~ cm_height, data = pset_data)
summary(model_2)
```

```
##
## Call:
## lm(formula = earnings ~ cm_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
## cm_height     278.61      19.88  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
```

E.i. What is the estimated slope of the regression?

The slope is 278.61. A one cm increase in height is associated with on average, a 278.61 dollar increase in annual earnings. Note that $278.61 = \frac{707.67}{2.54}$

E.ii. What is the estimated intercept?

The intercept is -512.73. Unlike the slope, changing the units of height does not change the intercept.

E.iii. What is the R^2?

The R^2 is 0.011. (rounding up)

E.iv. What is the standard error of the regression?

The standard error of the regression is 26780. Recall, that the error term ε_i is not observed, but we can estimate the residual $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$. The SER can be calculated using:

$$SER = \frac{SSR}{n-k}$$

Where $SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, n is the sample size, and k is the number of parameters in the regression. (This adjustment in the denominator i.e $n - k$, rather than n, accounts for the fact that there is a downward bias being introduced when > 1 parameter is estimated.)

The following code will replicate the output from the `R` regression table- there is a small difference due to rounding.

```
sqrt(sum(model_2$residuals^2) / (17870 - 2))
```

```
## [1] 26777.24
```

F. Run a regression of Earnings on Height, using data for female workers only.

We will use the `subset` function directly in the `lm` `data=` argument. Note: the variable sex is coded as 1=male, 0=female. It would have been nicer if instead of calling it sex, the variable name was male. This makes the inference that 1=male immediate.

```
model_3 <- lm(earnings ~ height, data = subset(pset_data, sex == 0))
summary(model_3)
```

```
##
## 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
```

F.i. What is the estimated slope?

The estimated slope is 511.2.

F.ii. A randomly selected woman is 1 inch taller than the average woman in the sample. Would you predict her earnings to be higher or lower than the average earnings for women in the sample? By how much?

First we need the average height for female workers in this sample. We estimate it using the following code.

```r
mean(pset_data$height[pset_data$sex == 0])
```

```
## [1] 64.49278
```

We can see that the average height for a female worker in the sample is 64.5. The predicted earnings for the average female worker in the sample is

$$12650.9 + (511.2 \times 64.49278) = 45619.61$$

You can verify that this is the *average earnings for female workers* in the sample. The intuition is that both *average height* and *average earnings*, both lie on the estimated regression line.

```r
# there is a minor difference due to rounding.
mean(pset_data$earnings[pset_data$sex == 0])
```

```
## [1] 45621
```

And for a female worker who is 1 inch taller than the average,

$$12650.9 + (511.2 \times 65.5) = 46134.5$$

A female worker who is 1 inch taller than the *average* female worker, is predicted to earn 511.2 dollars more per year.

G. Repeat (F) for male workers.

Follows from F, step-by-step.

```r
model_4 <- lm(earnings ~ height, data = subset(pset_data, sex == 1))
summary(model_4)
```

```
##
## 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
```

The estimated slope is 1306.9. The average height of male workers in the sample is 70 inches.

```
mean(pset_data$height[pset_data$sex == 1])
```

```
## [1] 70.08409
```

The predicted earnings for a male worker of average height is

$$-43130.3 + (1306.9 \times 70.1) = 48483.39$$

And for a male worker who is 1 inch taller than average,

$$-43130.3 + (1306.9 \times 71.1) = 49790.29$$

A male worker who is 1 inch taller than *average*, is predicted to earn 1306.9 dollars more per year.

H. Do you think that height is uncorrelated with other factors that cause earning? That is, do you think that the regression error term, say Ui, has a conditional mean of zero, given Height(Xi)?

Height (X_i) is correlated with other factors that influence earnings (u_i). One example of another factor is nutritional inputs in childhood. Height at adulthood is correlated with these nutritional inputs *AND* better nutrition can directly affect earnings (e.g. by improving attention span in school leading to improved cognitive skills that are known to increase earnings)