Econometric Methods

Solutions to empirical exercise 5.3, Chapter 5 S&W

Zaeen de Souza

14/07/2021

Contents

Background: Empirical Exercise 5.3

These are the solutions to **E5.3** 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 birthweight_smoking.
- 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("birthweight_smoking.xlsx")
```

Exercise E5.3

A. In the sample:

A.i. What is the average value of Birthweight for all mothers?

The average birthweight for all mothers in the sample is 3382.934.

```
mean(pset_data$birthweight, na.rm = T)
```

```
## [1] 3382.934
```

A.ii. For mothers who smoke?

The average birthweight for all mothers who smoke is 3178.832.

```
mean(pset_data$birthweight[pset_data$smoker == 1], na.rm = T)
```

```
## [1] 3178.832
```

A.iii. For mothers who do not smoke?

The average birthweight for all mothers who smoke is 3432.06.

```
mean(pset_data$birthweight[pset_data$smoker == 0], na.rm = T)
```

```
## [1] 3432.06
```

B

B.i. Use the data in the sample to estimate the difference in average birth weight for smoking and nonsmoking mothers.

The difference in birthweight between these groups is $3432.060 - 3178.832 = 253.228$. This version of the t=test syntax essentially tells R to run the test using the variable `smoker` as a grouping variable. Note the use of the ~ to indicate 'by group'.

```
# t-test
test <- t.test(birthweight ~ smoker, data = pset_data, var.equal = T)
test
```

```
##
##  Two Sample t-test
##
## data:  birthweight by smoker
## t = 9.3957, df = 2998, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   200.3831 306.0736
## sample estimates:
## mean in group 0 mean in group 1
##        3432.060        3178.832
```

We can see that the p-value is < 0.01 and as such, we can conclude that the difference in birthweight between these two groups is significant. Note, that here, the binary variable `smoker` takes the value 1 for smokers, and 0 for non-smokers.

Note, as well, that the mean birthweight in each group in the t-test output, matches exactly with our estimates from section `A.i` and `A.ii`. To replicate the t-test by hand, follow the steps in Problem set 1 in question `B.iii`.

B.ii. What is the standard error for the estimated difference in (i)?

We can recover the standard error of the estimated difference by accessing the the different items stored in the object we created called `test`. Note, that the the formula for the hand calculation is given in the last two problem sets—it should be straightforward to replicate here too. Here, we show how to extract it from an existing `R` function.

```
test$stderr
```

```
## [1] 26.95149
```

B.iii. Construct a 95% confidence interval for the difference in the average birth weight for smoking and nonsmoking mothers.

The confidence interval is given as part of the t-test output, and is $[200.3831, \ 306.0736]$. To manually calculate it we can:

$$253.228 \pm (10.96 \times 26.95149) = [200.4031, 306.0529]$$

Similar to the standard error, the confidence interval of the estimated difference is also part of the t-test output that `R` provides. We can access it from the stored object using the following:

```
test$conf.int
```

```
## [1] 200.3831 306.0736
## attr(,"conf.level")
## [1] 0.95
```

It also provides the level, chosen by the user in the `level=` argument of the t-test function. The default is 0.95 (95%). The minor difference in the hand calculation and the `R` output, is due to rounding.

C. Run a regression of Birthweight on the binary variable Smoker.

We want to estimate the following linear model using OLS:

$$Birthweight_i = \alpha + Smoker_i + \varepsilon_i$$

Where smoker takes the value 1 if mother i smoked and 0, if she didn't.

```
model_1 <- lm(birthweight ~ smoker, data = pset_data)
summary(model_1)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker, data = pset_data)
##
## Residuals:
##       Min       1Q   Median       3Q      Max
## -3007.06  -313.06    26.94   366.94  2322.94
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3432.06      11.87 289.115   <2e-16 ***
## smoker        -253.23      26.95  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.7 on 2998 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02828
## F-statistic: 88.28 on 1 and 2998 DF,  p-value: < 2.2e-16
```

C.i. Explain how the estimated slope and intercept are related to your answers in parts (a) and (b).

C.ii. Explain how the $SE(\hat{\beta}_1)$ is related to your answer in b(ii).

C.iii. Construct a 95% confidence interval for the effect of smoking on birth weight.

D. Do you think smoking is uncorrelated with other factors that cause low birth weight? That is, do you think that the regression error term, say ui, has a conditional mean of zero, given Smoking (Xi)? (You will investigate this further in Birthweight and Smoking exercises in later chapters.)