Econometric Methods: Solutions to Empirical Exercise 7.1 Chapter 7: Hypothesis Tests and Confidence Intervals in Multiple Regression Stock & Watson, 3^{rd} Edition

Zaeen de Souza [*] Deepti Goel[†]

Azim Premji University
22 February 2022

[*]

[†]Solution key prepared jointly by Zaeen and Deepti. R code and presentation in Rmarkdown by Zaeen.

Contents

Background: Empirical Exercise 7.1

These are the solutions to **E7.1** from **Chapter 7** 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")
```

E7.1 Problem Context

On the text website, www.pearsonglobaleditions.com/Stock_Watson, you will find the data file Birthweight_Smoking, which contains data for a random sample of babies born in Pennsylvania in 1989. The data include the baby's birth weight together with various characteristics of the mother, including whether she smoked during the pregnancy. A detailed description is given in Birthweight_Smoking_Description, also available on the website. In this exercise you will investigate the relationship between birth weight and smoking during pregnancy.

Exercise E7.1

Use the Birthweight_Smoking data set introduced in Empirical Exercise E5.3 to answer the following questions. To begin, run three regressions:

- (1) Birthweight on smoker
- (2) Birthhweight on smoker, alcohol, and nprevist
- (3) Birthweight on smoker, alcohol, nprevist, and unmarried

Running the three regressions in R (And using the table code from the last problem set (problem-set-4), we have:

```
model_1 <- lm(birthweight ~ smoker, data = pset_data)
model_2 <- lm(birthweight ~ smoker + alcohol + nprevist, data = pset_data)
model_3 <- lm(birthweight ~ smoker + alcohol + nprevist + unmarried, data = pset_data)
stargazer(model_1, model_2, model_3,
title = "Exercise E7.1",
header = F,
font.size = "small",
df = F,
digits = 3)
```

Table 1: Exercise E7.1

	Dependent variable:		
	birthweight		
	(1)	(2)	(3)
smoker	−253.228***	−217.580***	−175.377***
	(26.951)	(26.680)	(27.099)
alcohol		−30.491	−21.083
		(76.234)	(75.607)
nprevist		34.070***	29.603***
		(2.855)	(2.898)
unmarried			−187.133***
			(26.007)
Constant	3,432.060***	3,051.249***	3,134.400***
	(11.871)	(34.016)	(35.656)
Observations	3,000	3,000	3,000
R^2	0.029	0.073	0.089
Adjusted R^2	0.028	0.072	0.087
Residual Std. Error	583.730	570.471	565.698
F Statistic	88.279***	78.470***	72.793***
Note:			*p<0.1; **p<0.05; ***p<0.01

a. What is the value of the estimated effect of smoking on birth weight in each of the regressions?

Table 3 shows the estimated coefficients from each of the three regressions. In column 1, the estimated effect of mothers smoking on child's birthweight is -253.22, which implies that the children of mothers who smoked, were on average, -253.22 grams lighter than the children of mothers who didn't smoke. In column 2, where we introduce some controls, this coefficient is smaller in magnitude, and is -217.58. In column 3 with still more controls, we see that the coefficient is substantially smaller in magnitude, at around -175.377.

b. Construct a 95% confidence interval for the effect of smoking on birth weight, using each of the regressions.

Using R code, we have:

```
confint.lm(model_1, "smoker")
```

```
##             2.5 %    97.5 %
## smoker -306.0736 -200.3831
```

```
confint.lm(model_2, "smoker")
```

```
##             2.5 %    97.5 %
## smoker -269.8923 -165.2679
```

```
confint.lm(model_3, "smoker")
```

```
##             2.5 %    97.5 %
## smoker -228.5109 -122.2429
```

c. Does the coefficient on Smoker in regression (1) suffer from omitted variable bias? Explain.

Yes it does. Recall, that omitted variable bias occurs when there is some variable that is correlated with both our outcome of interest (birthweight), as well as our regressor of interest (smoker). In this case, alcohol, unmarried and number of pre-natal care visits are all potential controls, and as such, regression 1 suffers from omitted variable bias.

d. Does the coefficient on Smoker in regression (2) suffer from omitted variable bias? Explain.

Yes, regression 2 suffers from it as well. Marriage status can determine birthweight, if we assume that unmarried mothers are possibly younger, and on average, are less prepared for childbirth. It is also likely to be correlated with smoking status via prevailing social norms.

e. Consider the coefficient on Unmarried in regression (3).

e i. Construct a 95% confidence interval for the coefficient.

```
confint(model_3, "unmarried")
```

```
##                2.5 %    97.5 %
## unmarried -238.1276 -136.1389
```

e ii. Is the coefficient statistically significant? Explain.

Based on the confidence interval, and assuming we are testing the hypothesis that $\beta_4 = 0$, we could conclude that it is significant since the interval doesn't pass through 0.

e iii. Is the magnitude of the coefficient large? Explain.

The coefficient is economically significant—on average, the children of unmarried mothers weighed 187 grams less than the children of married mothers, even after controlling for alcohol intake, number of pre-natalcase checks, and smoking habits. More intuitively, children of unmarried mothers weighed 5% less than the average child of a non-smoking mother.

e iv. A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree? (Hint: Review the discussion of control variables in Section 7.5. Discuss some of the various factors that Unmarried may be controlling for and how this affects the interpretation of its coefficient.)

The advocacy group is jumping to conclusions. Unmarried mother could be a stand in for other life style choices that affect birth weight of the mother such as teenage pregnancy, risky lifestyle choices, and low income. It could be that addressing these underlying issues may result in 'unmarried' status having no effect on birth weight of its own.

f. Consider the various other control variables in the data set. Which do you think should be included in the regression? Using a table like Table 7.1, examine the robustness of the confidence interval you constructed in (b). What is a reasonable 95% confidence interval for the effect of smoking on birth weight?

In order to examine this, we could estimate the following linear model using OLS,

$$Y_i = \alpha + \beta_1 Smoker_i + \beta_2 nprevisit_i + \beta_3 Alcohol_i + \beta_4 Age_i + \beta_5 Edu_i + \beta_6 Unmarried_i + \varepsilon_i$$

Table 2 shows the result of this regression—covariates are added one at a time.

```
model_1 <- lm(birthweight ~ smoker, data = pset_data)
model_2 <- lm(birthweight ~ smoker + nprevist, data = pset_data)
model_3 <- lm(birthweight ~ smoker + nprevist + alcohol, data = pset_data)
model_4 <- lm(birthweight ~ smoker + nprevist + alcohol + age, data = pset_data)
model_5 <- lm(birthweight ~ smoker + nprevist + alcohol + age + educ, data = pset_data)
model_6 <- lm(birthweight ~ smoker + nprevist + alcohol + age + educ + unmarried, data = pset_
```

A reasonable interval for this correlation would be based on the regression in column 6, which included a full set of controls. The interval is given by -

```
confint(model_6, "smoker")
```

```
##             2.5 %    97.5 %
## smoker -230.8283 -123.0894
```

Table 2: Exercise E7.1

	Dependent variable:					
	birthweight					
	(1)	(2)	(3)	(4)	(5)	(6)
smoker	−253.228***	−218.829***	−217.580***	−211.518***	−206.507***	−176.959***
	(26.951)	(26.492)	(26.680)	(26.923)	(27.367)	(27.474)
nprevist		34.104***	34.070***	33.426***	32.979***	29.775***
		(2.853)	(2.855)	(2.881)	(2.914)	(2.926)
alcohol			−30.491	−38.191	−39.512	−14.758
			(76.234)	(76.354)	(76.365)	(75.839)
age				3.277*	2.360	−2.494
				(1.984)	(2.178)	(2.269)
educ					5.644	0.238
					(5.532)	(5.542)
unmarried						−199.319***
						(28.396)
Constant	3,432.060***	3,050.527***	3,051.249***	2,969.189***	2,924.963***	3,199.426***
	(11.871)	(33.963)	(34.016)	(60.204)	(74.185)	(83.337)
Observations	3,000	3,000	3,000	3,000	3,000	3,000
R^2	0.029	0.073	0.073	0.074	0.074	0.089
Adjusted R^2	0.028	0.072	0.072	0.072	0.072	0.087
Residual Std. Error	583.730	570.391	570.471	570.306	570.302	565.760
F Statistic	88.279***	117.658***	78.470***	59.568***	47.864***	48.741***
Note:						*p<0.1; **p<0.05; ***p<0.01