Econometric Methods: Solutions to Empirical Exercise 6.1 Chapter 6: Multiple Regression Stock & Watson, 3^{rd} Edition

Zaeen de Souza [*] Deepti Goel[†]

Azim Premji University
09 February 2022

[*]

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

Contents

Background: Empirical Exercise 6.1

These are the solutions to **E6.1** from **Chapter 6** 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 6 Multiple Regression")

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

E6.1 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 E6.1

a. Regress Birthweight on Smoker. What is the estimated effect of smoking on birth weight?

We will estimate,

$$Birthweight_i = \alpha + \beta_1 Smoker_i + \varepsilon_i$$

```
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
```

The regression coefficient on smoker, -253.23, reveals that the average birthweight for mothers who smoke is 253.23 grams lower than the average birthweight for mothers who do not smoke.

b. Regress Birthweight on Smoker, Alcohol, and Nprevist.

```
model_2 <- lm(birthweight ~ smoker + alcohol + nprevist, data = pset_data)
summary(model_2)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + nprevist, data = pset_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2733.53  -307.57    21.42   358.09  2192.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3051.249     34.016  89.701  < 2e-16 ***
## smoker      -217.580     26.680  -8.155 5.07e-16 ***
## alcohol      -30.491     76.234  -0.400    0.689
## nprevist      34.070      2.855  11.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.5 on 2996 degrees of freedom
## Multiple R-squared:  0.07285,    Adjusted R-squared:  0.07192
## F-statistic: 78.47 on 3 and 2996 DF,  p-value: < 2.2e-16
```

b.i Using the two conditions in Key Concept 6.1, explain why the exclusion of Alcohol and Nprevist could lead to omitted variable bias in the regression estimated in (a).

When `alcohol` and `nprevisit` are included as controls, the coefficient on smoker increases from -253.23 to -217.58 (magnitude is reduced). This is because adding controls to a regression boils down to estimating the association between birthweight and smoking from a comparison of smokers and non-smokers who have similar levels of alcohol consumption and prenatal visits. Such an association comes closer to the 'causal' effect of smoking.

Now let us understand the drop in magnitude of the coefficient in terms of the omitted variables bias formula: take alcohol first: birthweight and alcohol are negative correlated, and alcohol and smoking are positively correlating, so in the absence of alcohol as an explicit regressor the coefficient in the short regression should suffer from negative bias, i.e., it should be a larger negative number. Next, let us look at pre natal visits. Birthweight and prenatal visits are positively correlated, and prenatal visits and smoking are negatively correlated, so in the absence of nprevist as an explicit regressor the coefficient in the short regression should suffer from negative bias i.e. it should be a larger negative number.

b.ii Is the estimated effect of smoking on birthweight substantially different from the regressionthat excludes Alcohol and Nprevist? Does the regression in (a) seem to suffer from omitted variable bias?

As discussed in b.i, the regression in (a) does suffer from omitted variables bias when we are thinking of using it to estimate the causal effect of smoking on birthweight. Note, however, that the regression in (a) gives a consistent estimator of $E(Y|Smoker)$.

b iii. Jane smoked during her pregnancy, did not drink alcohol, and had 8 prenatal care visits. Use the regression to predict the birth weight of Jane's child.

We can predict the birth weight of her child as follows:

$$\hat{Y} = 3051.249 + (-217.580 \times 1) + (-30.491 \times 0) + (34.070 \times 8)$$

Which gives us a predicted birth weight of 3106.229 grams.

b iv. Compute R^2 and $\bar{R}^2$ Why are they so similar?

Recall,

$$R^2 = \frac{ESS}{TSS} = 1 - (\frac{SSR}{TSS})$$

Where,

- The Explained Sum of Squares (ESS) is defined as: $ESS = \sum_{i=1}^{n}(\hat{Y} - \bar{Y})^2$
- The Total Sum of Square (TSS) is defined as: $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- The Sum of Squared Residuals (SSR) is defined as: $SSR = \sum_{i=1}^{n} \hat{u}_i$. Recall, that $\hat{u}_i = Y_i - \hat{Y}_i$ are the estimated residuals from the regression.

To build intuition, consider the following explanation, in words: The R^2 is the explained sum of squares (ESS) divided by the total sum of squares (TSS)—it is the proportion of variance in some outcome Y that is explained by some $\mathbf{X}$, where $\mathbf{X}$ could be a matrix of any dimension (i.e $\mathbf{X}^{m \times p}$.

The adjusted-R^2 i.e $\bar{R}^2$, is 1 minus the ratio of the variance of the OLS residuals (along with a degrees of freedom correction) to the variance of Y. It is calculated as follows,

$$\bar{R}^2 = 1 - (\frac{n-1}{n-k-1}) \cdot \frac{SSR}{TSS}$$

Below, we have the hand calculation for the R^2

```
# the predict() function calculates the predicted Y from the regression we ran eariler.
pset_data$yhat <- predict(model_2)
ssr <- sum((pset_data$yhat - pset_data$birthweight)^2)
tss <- sum((pset_data$birthweight - mean(pset_data$birthweight))^2)
rsq <- 1 - ssr/tss
rsq
```

```
## [1] 0.0728503
```

Similarly, for the $\bar{R}^2$,

```
1-(1-summary(model_2)$r.squared)*((nobs(model_2)-1)/(nobs(model_2)-1-3))
```

```
## [1] 0.07192191
```

Doing this via code and the `object` already saved by `R`:

```
summary(model_2)$r.squared
```

```
## [1] 0.0728503
```

```
summary(model_2)$adj.r.squared
```

```
## [1] 0.07192191
```

The two values are similar for the following reason: n is sufficiently large ($n = 3000$) and given that there are only a few regressors (in this case, 3) the degree of freedom adjustment is also trivially small, and as such, the R^2 and $\bar{R}^2$ are qualitatively the same.

c. Estimate the coefficient on Smoking for the multiple regression model in (b), using the three-step process in Appendix (6.3) (the Frisch-Waugh theorem). Verify that the three-step process yields the same estimated coefficient for Smoking as that obtained in (b).

The three steps to estimate $\widehat{\beta}_1 Smoker$ are as follows:

1. Regress X_1 on X_2, $X_3 \ldots$, X_k and obtain $\widetilde{X}_1$ which are the residuals from this regression.
2. Regress Y on X_2, $X_3 \ldots$, X_k and obtain $\widetilde{Y}$ which are the residuals from the second regression.
3. Regress $\widetilde{Y}$ on $\widetilde{X}_1$

Using R to implement this, we have:

```
# step 1
step_1 <- lm(smoker ~ alcohol + nprevist, data = pset_data)
# storing the estimated residuals from this regression
pset_data$xtilde <- residuals(step_1)
# step 2
step_2 <- lm(birthweight ~ alcohol + nprevist, data = pset_data)
# storing the estimated residuals from this regression
pset_data$ytilde <- residuals(step_2)
# step 3
step_3 <- lm(ytilde ~ xtilde, data = pset_data)
# summary(step_3) if you want to see the summary in your console
```

After doing this, one can either do `summary(step_3)` as mentioned above, and look at the coefficient on `xtilde` and verify that it is indeed the same as the coefficient on smoker in question b. (model_2). For ease of comparison, table 1 shows a side-by-side comparison of these two regressions - the coefficient xtilde in column 2 matches the coefficient on smoker in column 1.

Note as well, that while the coefficients match, the test statistic and standard error will not unless we use the correct degrees of freedom - R does not know that we have estimated this in the way we

did, and as such, has falsely inflated the degrees of freedom. If this adjustment is made, the error and test statistics will match too.

As an aside - this is the code to make a latex regression table in R:

```
library(stargazer)
stargazer(model_2, step_3, digits = 4,
df = F,
title = "Frisch-Waugh-Lovell: An Example",
out.header = F,
header = F,
dep.var.labels = c("Birthweight", "ytilde"),
column.labels = c("Multiple Regression",
"FWL-Bivariate Regression"))
```

Table 1: Frisch-Waugh-Lovell: An Example

	Dependent variable:	
	Birthweight Multiple Regression	ytilde FWL-Bivariate Regression
	(1)	(2)
smoker	-217.5801^{***}	
	(26.6796)	
alcohol	-30.4913	
	(76.2340)	
nprevist	34.0699^{***}	
	(2.8550)	
xtilde		-217.5801^{***}
		(26.6707)
Constant	$3{,}051.2490^{***}$	0.0000
	(34.0160)	(10.4119)
Observations	3,000	3,000
R^2	0.0729	0.0217
Adjusted R^2	0.0719	0.0214
Residual Std. Error	570.4708	570.2805
F Statistic	78.4697^{***}	66.5533^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

d. An alternative way to control for prenatal visits is to use the binary variables `Tripre0` through `Tripre3`. Regress `Birthweight` on `Smoker`, `Alcohol`, `Tripre0`, `Tripre2`, and `Tripre3`.

We will estimate the following linear model using OLS,

$$Birthweight_i = \alpha + \beta_1 Smoker_i + \beta_2 Alcohol_i + \beta_3 Tripre0_i + \beta_4 Tripre2_i + \beta_5 Tripre3_i + \varepsilon_i$$

Where i indexes the individual, The variables are defined already. We use the following code:

```
model_3 <- lm(birthweight ~ smoker + alcohol + tripre0 +
tripre2 + tripre3, data = pset_data)
summary(model_3)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + tripre0 + tripre2 +
##     tripre3, data = pset_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3029.55  -307.55    31.35   372.45  2401.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3454.55      12.65 273.077  < 2e-16 ***
## smoker       -228.85      27.16  -8.424  < 2e-16 ***
## alcohol       -15.10      77.54  -0.195 0.845613
## tripre0      -697.97     106.88  -6.531 7.66e-11 ***
## tripre2      -100.84      29.62  -3.404 0.000672 ***
## tripre3      -136.96      59.58  -2.299 0.021595 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 578.7 on 2994 degrees of freedom
## Multiple R-squared:  0.04647,    Adjusted R-squared:  0.04487
## F-statistic: 29.18 on 5 and 2994 DF,  p-value: < 2.2e-16
```

d i. Why is Tripre1 excluded from the regression? What would happen if you included it in the regression?

If we included it then we have Tripe0 + Tripe1 + Tripe2 + Tripe3 = Intercept'. Thus, one of the variables, namely, the intercept (or constant term) can be explained as a linear combination of the other included variables. This is called perfect multicollinearity. Intuitively, since once we know any four variables, we can compute the fifth variable, the information in one of them is redundant. Thus, we need to drop any one of these five variables and then estimate the model. Typically, the intercept is retained and one of the other dummies is dropped.

An example:

```
model_4 <- lm(birthweight ~ smoker + alcohol + tripre0 + tripre1 +
tripre2 + tripre3, data = pset_data)
summary(model_4)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + tripre0 + tripre1 +
##     tripre2 + tripre3, data = pset_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3029.55  -307.55    31.35   372.45  2401.29
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3317.59      59.00  56.231  < 2e-16 ***
## smoker       -228.85      27.16  -8.424  < 2e-16 ***
## alcohol       -15.10      77.54  -0.195   0.8456
## tripre0      -561.01     120.88  -4.641 3.61e-06 ***
## tripre1       136.96      59.58   2.299   0.0216 *
## tripre2        36.12      64.17   0.563   0.5736
## tripre3           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 578.7 on 2994 degrees of freedom
## Multiple R-squared:  0.04647,    Adjusted R-squared:  0.04487
## F-statistic: 29.18 on 5 and 2994 DF,  p-value: < 2.2e-16
```

Here R automatically drops one of the dummy variables - in this case tripre3, and this used as a "base category".

d ii. The estimated coefficient on Tripre0 is large and negative. What does this coefficient measure? Interpret its value.

Looking at the regression output in question d i, we can see that after controlling for alcohol consumption and smoking status, the children of mothers who had no pre-natal care visits, on average, weigh 697.97 grams less than the children of mothers who had at their first pre-natal care visit in the first trimester of their pregnancy.

d iii. Interpret the value of the estimated coefficients on Tripre2 and Tripre3.

Using the regression output in question d i, we can interpret the coefficients for the dummy variables in the following way. The children of mothers who had their first pre-natal care visit in the 2nd trimester (Tripre2=1) on average, weighed a 100.84 grams less, compared to children of mothers who had their first pre-natal care visit during their first trimester of pregnancy. The children of mothers who had their first pre-natal care visit in the 3rd trimester (Tripre3=1) on average, weighed a 136.96 grams less, compared to children of mothers who had their first pre-natal care visit during

their first trimester of pregnancy.

d iv. Does the regression in (d) explain a larger fraction of the variance in birth weight than the regression in (b)?

The regression in (b) where the $R^2 = 0.073$, explains more of the variation in birth weight than the regression in (d), where the $R^2 = 0.046$. This can be interpreted in the following way. It could be the case that the timing of the first prenatal visit is not as important in determining birth weight than the overall number of prenatal checks that a mother received.Note as well that including more covariates in the regression tends to increase the R^2. Looking at the $\bar{R}^2$, we can see that even here, regression (b) has an $\bar{R}^2 = 0.072$ and regression (d) has an $\bar{R}^2 = 0.045$.