Assignment

The assignment was to read chapter 3 and hand in answers to the following problems at the end of the chapter: C3.1 – C3.8.

C3.1

A problem of interest to health officials (and others) is to determine the effects of smoking during pregnancy on infant health. One measure of infant health is birth weight; a birth weight that is too low can put an infant at risk for contracting various illnesses. Since factors other than cigarette smoking that affect birth weight are likely to be correlated with smoking, we should take those factors into account. For example, higher income generally results in access to better prenatal care, as well as better nutrition for the mother. An equation that recognizes this is

\[ \text{bwght} = \beta_0 + \beta_1 \text{cigs} + \beta_2 \text{faminc} + u \]

i

What is the most likely sign for \( \beta_2 \)?
If higher income increases access to prenatal care, and prenatal care increases birth weight, then I expect \( \text{faminc} \) to be positive.

ii

Do you think \( \text{cigs} \) and \( \text{faminc} \) are likely to be correlated? Explain why the correlation might be positive or negative.
Cigarette smoking and income might well be correlated. I expect that they would be negatively correlated.
Now, estimate the equation with and without faminc, using the data in BWGHT.RAW. Report the results in equation form, including the sample size and R-squared. Discuss your results, focusing on whether adding faminc substantially changes the estimated effect of cigs and bwght.

First, the regression with faminc included:

\[ \hat{bwght} = 116.97 - 0.46cigs + 0.09faminc \]
\[ N = 1,388 \quad R - squared = 0.03. \]

Now, the regression with faminc omitted:

\[ \hat{bwght} = 119.77 - 0.51cigs \]
\[ N = 1,388 \quad R - squared = 0.02. \]

The omission of the faminc variable increases the magnitude of the effect of cigs. This is to be expected if faminc and cigs are negatively related. The effect of faminc on bwght is positive, and the effect of faminc on cigs is negative, meaning that the net effect of omitting faminc from the model is to make the effect of cigs on bwght less (i.e. more negative).

C3.2

Use the data in HPRICE1.RAW to estimate the model

\[ price = \beta_0 + \beta_1sqrft + \beta_2bdrms + u \]

where price is the house price measured in thousands of dollars.

i

Write out the results in equation form.

\[ \hat{price} = -19.31 + 0.13sqrft + 15.20bdrms \]
\[ N = 88 \quad R - squared = 0.63 \]

ii

What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

One more bedroom is estimated to increase the sales price by $15,200.
What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).
Adding a bedroom without increasing the size of the house at all results in an increase in price of $15,200. Doing so essentially means that you would be adding a bedroom (which takes up some number of square feet) and subtracting that number of square feet from elsewhere in the house (so that you gained a bedroom without adding any square feet). If we now add a bedroom and 140 square feet to a house, we increase its predicted sales price by \(0.13 \times 140 + 15.20 \times 1 = 33.4\), or $33,400 (your numbers may differ due to rounding).

What percentage of the variation in price is explained by square footage and number of bedrooms?
Approximately 63% of the variation in price is explained by square footage and the number of bedrooms.

The first house in the sample has \(\text{sqrft} = 2,438\) and \(\text{bdrms} = 4\). Find the predicted selling price for this house from the OLS regression line.
The predicted selling price for such a house is \[358.43 = -19.31 + 0.13(2438) + 15.20(4),\]
or $358,430.

The actual selling price of the first house in the sample was $300,000 (so price = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?
If the actual selling price was $300,000 and the predicted sales price was $358,430, then the residual is the actual less the predicted value, or $300,000 - $358,430 = -$58,430. 
This suggestion (upon initial inspection) is that the buyers underpaid for the house (good deal!). But of course we must realize that there are lots of things that determine the price of a house: location (schools), lot size, number of bathrooms, garage, appliances, etc. So it might be the case that although the average house with 2,438 sq. ft. and 4 bdrms goes for $358,430, a house of the same size but in a neighborhood without the best schools, etc. might go for significantly less.
C3.3

The file CEOSAL2.RAW contains data on 177 chief executive officers and can be used to examine the effects of firm performance on CEO salary.

i

Estimate a model relating annual salary to firm sales and market value. Make the model the constant elasticity variety for both independent variables. Write the results out in equation form.

The “constant elasticity variety” means a model that is linear in elasticities. Elasticities are percentage changes. So a constant elasticity model would be:

\[ \log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 \log(mktval) + u. \]

When I estimate such a model, I get:

\[ \hat{\log(salary)} = 4.62 + 0.16 \log(sales) + 0.11 \log(mktval) \]

\[ N = 177 \quad R - squared = 0.30. \]

ii

Add profits to the model from part (i). Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?

We cannot add profits in logarithmic form because profits take on negative values. We cannot take the log of a negative value (it is undefined). When I run the new model I get:

\[ \hat{\log(salary)} = 4.69 + 0.16 \log(sales) + 0.10 \log(mktval) + 0.00 \text{profits} \]

\[ N = 177 \quad R - squared = 0.30. \]

I would NOT, in fact, say that these firm performance variables explain most of the variation in CEO salaries. The R-squared is approximately 0.3, meaning that 70% of the variation in log(salary) is unexplained. Profits seems to add very little to the model, suggesting that profits have very little influence on log(salary).

iii

Add the variable ceoten to the model in part (ii). What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?

When I add ceoten to the model, I get:

\[ \hat{\log(salary)} = 4.56 + 0.16 \log(sales) + 0.10 \log(mktval) + 0.00 \text{profits} + 0.01 \text{ceoten} \]

\[ N = 177 \quad R - squared = 0.32. \]

The implication is that when CEO tenure increases by one year, salary increases by 1%.
Find the sample correlation coefficient between the variables \(\log(mktval)\) and profits. Are these variables highly correlated? What does this say about the OLS estimators? The sample correlation coefficient is 0.78. \(\log(mktval)\) and profits are highly correlated. Profits and the log of market value move together, suggesting that estimating the independent effect of each on \(\log(salary)\) is difficult. The upshot here is that we should not be surprised that we do not find a large (independent) impact of both mktval and profits.

C3.4

Use the data in ATTEND.RAW for this exercise.

i

Obtain the minimum, maximum, and average values for the variables atndrte, priGPA, and ACT.

To obtain the min, max, and mean of values in Stata, simply use the `sum` function.

<table>
<thead>
<tr>
<th>variable</th>
<th>min</th>
<th>max</th>
<th>mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>atndrte</td>
<td>6.25</td>
<td>200</td>
<td>81.71</td>
</tr>
<tr>
<td>priGPA</td>
<td>0.86</td>
<td>3.93</td>
<td>2.59</td>
</tr>
<tr>
<td>ACT</td>
<td>13</td>
<td>32</td>
<td>22.51</td>
</tr>
</tbody>
</table>

ii

Estimate the model

\[
\text{atndrte} = \beta_0 + \beta_1 \text{priGPA} + \beta_2 \text{ACT} + u,
\]

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

The estimated equation is:

\[
\hat{\text{atndrte}} = 75.70 + 17.26 \text{priGPA} - 1.72 \text{ACT}
\]

\(N = 680 \quad R - squared = 0.29.\)

The intercept of 75.70 is the predicted percent of classes attended for a student with 0 cumulative GPA prior to the current term and an ACT score of 0. I would not call this particular meaning “useful.” The intercept is useful, but its interpretation is not.
Discuss the estimated slope coefficients. Are there any surprises?
One extra GPA point (priGPA) is predicted to increase the percentage of classes attended by about 17 percent. This makes a reasonable amount of sense — increasing GPA increases attendance. One extra point on the ACT exam, on the other hand, is predicted to decrease attendance. This is unexpected. You could interpret the results in a way that makes some sense, but it’s not my first choice of story: maybe an increase in ACT score indicates an increase in ability, and with increased ability it becomes less necessary to attend classes. I don’t particularly like this story — I would call it a reach at best — but it is consistent with the data.

What is the predicted atndrte if priGPA = 3.65 and ACT = 20? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?
When priGPA is 3.65 and ACT is 20, atndrte is predicted to be:

\[ 104.30 = 75.7 + 17.26 \times (3.65) - 1.72 \times (20). \]

A student with a GPA of 3.65 and an ACT of 20 would seem to be a very good student. But no student attends more than 100% of classes! There is one student with these exact values (to find them in Stata, type `list if priGPA>3.64 & priGPA<3.66`. You’ll see that only two observations meet this criteria. One of them (observation number 569) has an ACT of exactly 20 (if you type `list if priGPA==3.65` you will not get any observations listed, due to rounding). This student actually attends 87.5% of classes.

If Student A has priGPA = 3.1 and ACT = 21 and Student B has priGPA = 2.1 and ACT = 26, what is the predicted difference in their attendance rates?
Student A:

\[ 93.09 = 75.7 + 17.26 \times (3.1) - 1.72 \times (21). \]

Student B:

\[ 67.23 = 75.7 + 17.26 \times (2.1) - 1.72 \times (26). \]

The difference in predicted attendance between Student A and Student B is 93.09 - 67.23 = 25.86%.

C3.5

Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partially out for Example 3.2. This first requires regressing educ on exper and tenure
and saving the residuals $\hat{r}_1$. Then, regress $\log(\text{wage})$ on $\hat{r}_1$. Compare the coefficient on $\hat{r}_1$ with the coefficient on educ in the regression of $\log(\text{wage})$ on educ, exper, and tenure.

Use WAGE1.RAW. The target of the “partialling out” procedure is obtaining the coefficient on educ in the regression:

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + u.$$  

When I run this regression in Stata I get:

$$\hat{\log(\text{wage})} = .28 + .09\text{educ} + .00\text{exper} + .02\text{tenure} \quad N = 526 \quad R - squared = 0.32.$$  

So we are trying to recover 0.09. This is the exact same procedure that we carried out in class. What we are doing is trying to find the effect of educ on $\log(\text{wage})$, controlling for exper and tenure. This effect is equal to the effect on $\log(\text{wage})$ of the portion of educ that is NOT explained by exper and tenure. First we need to construct a variable that is equal to the portion of educ that is not explained by exper and tenure. The easiest way to do that is to take the residual from the regression:

$$\text{educ} = \gamma_0 + \gamma_1 \text{exper} + \gamma_2 \text{tenure} + v.$$  

(I use $\gamma$ and $v$ rather than $\beta$ and $u$ to distinguish the coefficients from those in the main regression above).

When I perform this first-stage regression I get:

$$\hat{\text{educ}} = 13.57 - 0.07\text{exper} + 0.05\text{tenure} \quad N = 526 \quad R - squared = 0.10.$$  

To find the residuals in this regression I subtract $\hat{\text{educ}}$ from educ:

$$\hat{r}_1 = \text{educ} - \hat{\text{educ}}.$$  

When I then regress $\log(\text{wage})$ on $\hat{r}_1$ I get:

$$\hat{\log(\text{wage})} = 1.62 + 0.09r1 \quad N = 526 \quad R - squared = 0.21.$$  

C3.6  

Use the data set in WAGE2.RAW for this problem. As usual, be sure all of the following regressions contain an intercept.
i

Run a simple regression of IQ on educ to obtain the slope coefficient, say \( \tilde{\delta}_1 \).

Result:

\[
\hat{IQ} = 53.69 + 3.53\text{educ}
\]

\( N = 935 \quad R - squared = 0.27 \).

ii

Run the simple regression of log(wage) on educ, and obtain the slope coefficient, \( \tilde{\beta}_1 \).

Result:

\[
\hat{\log(\text{wage})} = 5.97 + 0.06\text{educ}
\]

\( N = 935 \quad R - squared = 0.10 \).

iii

Run the multiple regression of log(wage) on educ and IQ, and obtain the slope coefficients, \( \hat{\beta}_1 \) and \( \hat{\beta}_2 \), respectively.

Result:

\[
\hat{\log(\text{wage})} = 5.66 + 0.04\text{educ} + 0.01\text{IQ}
\]

\( N = 935 \quad R - squared = 0.13 \).

iv

Verify that \( \hat{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 \). When we take \( \hat{\beta}_1 = 0.04 \) (from the last regression), and add to it \( \hat{\beta}_2 = 0.01 \) (again from the last regression) times \( \tilde{\delta}_1 = 3.53 \) (from the very first regression), we get \( 0.04 + 0.01 \times 3.53 = 0.07 \). I guess my habit of rounding to two digits isn’t cutting the mustard here. OK, I’ll do it the long way: \( 0.0391199 + 0.0058631 	imes 3.533829 = 0.05983909 \). There, that worked.

The more important question is: what’s the takeaway? The best way to see is probably to look at the results this way instead:

\[
\hat{\beta}_1 = \hat{\beta}_1 - \hat{\beta}_2 \tilde{\delta}_1.
\]

The effect of educ on log(wage), controlling for IQ, is given by the effect of educ on log(wage) (without any controls), less the effect on log(wage) of of IQ, and the effect on log(wage) that is unassignable (the stuff that is essentially caused simultaneously by both educ and IQ). To do this, we want to subtract out what is contained in \( \hat{\beta}_2 \) (the effect of IQ on log(wage) that is independent of educ). But we have to weight this effect (\( \hat{\beta}_2 \)) by how much educ moves when IQ moves.
C3.7

Use the data in MEAP93.RAW to answer this question.

i

Estimate the model

$$\text{math10} = \beta_0 + \beta_1 \log(\text{expend}) + \beta_2 \ln\text{chprg} + u,$$

and report the results in the usual form, including the sample size and R-squared. Are the signs of the slope coefficients what you expected? Explain.

Result:

$$\hat{\text{math10}} = -20.36 + 6.23 \log(\text{expend}) - 0.30 \ln\text{chprg}$$

$$N = 408 \quad R - \text{squared} = 0.18.$$  

The sign of the coefficients are as expected: the percentage of students passing a math exam is increasing in expenditure per student and decreasing in the percentage of students who are in a school lunch program (presumably a subsidized lunch program).

ii

What do you make of the intercept in part(i)? In particular, does it make sense to set the two explanatory variables to zero?

The constant doesn’t have much of an interpretation. While it makes fine sense to set \(\ln\text{chprg}\) to 0 (we could have no students in the lunch program), it makes little sense to set \(\log(\text{expend})\) to 0. Doing so would require \(\text{expend}\) to be equal to 1: spending $1 per student. The minimum of \(\text{expend}\) in the data is $3,332.

iii

Now run the simple regression of math10 on \(\log(\text{expend})\), and compare the slope coefficient with the estimate obtained in part(i). Is the estimated spending effect now larger or smaller than in part(i)?

When I run the simple regression of \(\text{math10}\) on \(\log(\text{expend})\) I get:

$$\hat{\text{math10}} = -69.34 + 11.16 \log(\text{expend})$$

$$N = 408 \quad R - \text{squared} = 0.03.$$  

The magnitude of the slope coefficient has gotten larger. It was previously 6.23 and is now 11.16. This speaks to a negative correlation between \(\log(\text{expend})\) and \(\ln\text{chprg}\).
iv

Find the correlation between \( \text{lexpend} = \log(\text{expend}) \) and \( \lnchprg \). Does its sign make sense to you?

Since \( \lnchprg \) is a subsidized lunch program, yes. Although running a lunch program would cause higher expenditure, all else equal, I expect that schools with a lower percentage of children enrolled in a subsidy program are likely to spend more per student in total, i.e. more on other things throughout the school. Bottom line: wealthier jurisdictions are likely to spend more per student.

v

Use part(iv) to explain your findings in part(iii).

I answered this in the previous part — the fact that the coefficient on \( \log(\text{expend}) \) got larger when we excluded \( \lnchprg \) indicates, of course, that the inclusion of \( \lnchprg \) suppressed the coefficient on \( \log(\text{expend}) \). What would lead this to happen? To see, combine two facts: (1) when \( \lnchprg \) increases, \( \text{math10} \) decreases; (2) when \( \text{lexpend} \) increases, \( \lnchprg \) decreases. Therefore, when \( \text{lexpend} \) increases, \( \lnchprg \) decreases, which causes \( \text{math10} \) to go ... up.

C3.8

Use the data in DISCRIM.RAW to answer this question. These are ZIP code-level data on prices for various items at fast-food restaurants, along with characteristics of the zip code population, in New Jersey and Pennsylvania. The idea is to see whether fast-food restaurants charge higher prices in areas with a larger concentration of blacks.

i

Find the average values of \( \text{prpb1ck} \) and \( \text{income} \) in the sample, along with their standard deviations. What are the units of measurement of \( \text{prpb1ck} \) and \( \text{income} \)?

Once again, this can be done with the \texttt{sum} function in Stata. Result:

<table>
<thead>
<tr>
<th>variable</th>
<th>mean</th>
<th>std. dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>prpb1ck</td>
<td>0.11</td>
<td>0.18</td>
</tr>
<tr>
<td>income</td>
<td>47053.78</td>
<td>13179.29</td>
</tr>
</tbody>
</table>

ii

Consider a model to explain the price of soda, \( \text{psoda} \), in terms of the proportion of the population that is black and median income:

\[
\text{psoda} = \beta_0 + \beta_1 \text{prpb1ck} + \beta_2 \text{income} + u.
\]
Estimate this model by OLS and report the results in equation form, including the sample size and R-squared. (Do not use scientific notation when reporting the estimates.) Interpret the coefficient on prpblck. Do you think it is economically large?

Result:

\[ \hat{psoda} = 0.95 + 0.11prpblck + 0.00\text{income} \]

\[ N = 401 \quad R \text{ – squared} = 0.06 \]

The coefficient on prpblck is 0.1149882. The literal interpretation would be: when prpblck increases by 1, the price of a medium soda increases by 11 cents. The only problem is, the notion of increasing prpblck by 1 is not very meaningful. prpblck is the proportion of individuals in a zip code who are black cannot increase by 1 unless the proportion of individuals in a zip code starts out as 0. That is, the only zip code that can increase by 1 is a zip code that starts out with no individuals who are black, and then becomes a zip code that is made up only of individuals who are black. This is not a very useful marginal effect. In order to interpret the marginal effect more usefully, look at smaller (more realistically-sized) changes. For instance, an increase of 0.01 (an increase of 1 in the percentage of individuals who are black in a zip code) is predicted to increase the price of a medium soda by 0.1149882 \times 0.01 = 0.00114988, or approximately not at all. An increase of 0.10 (an increase of 10 in the percentage of individuals who are black in a zip code) is predicted to increase the price of a medium soda by 0.1149882 \times 0.10 = 0.0114988, or approximately a penny. Is this economically large? I do not think so, on its face. If, however, there are many medium sodas purchased, such an effect might be better expressed in terms of total expenditure, which might be large. A penny, multiplied by many tens of thousands of sodas in a particular zip code, starts to run into real money.

iii

Compare the estimate from part(ii) with the simple regression estimate from psoda on prpblck. Is the discrimination effect larger or smaller when you control for income?

Result of a simple regression of psoda on prpblck:

\[ \hat{psoda} = 1.04 + 0.06prpblck \]

\[ N = 401 \quad R \text{ – squared} = 0.02 \]

The discrimination effect is estimated to be significantly smaller when income is excluded from the regression.

iv

A model with a constant price elasticity with respect to income may be more appropriate. Report estimates of the model

\[ \log(psoda) = \beta_0 + \beta_1prpblck + \beta_2\log(income) + u. \]
If prpblk increases by 0.20 (20 percentage points), what is the estimated percentage change in psoda?
The result of the model is:
\[
\hat{psoda} = -0.79 + 0.12prpblk + 0.08\log(income)
\]
\[N = 401 \quad R - squared = 0.07.\]

Notice that we do not need to create a \log(prpblk) variable. prpblk is already a proportion, meaning that we can already interpret the coefficient on prpblk in terms of percentage changes. If prpblk increases by 0.20, then the price of soda is predicted to increase by \[0.12 \times 0.20 = 0.024 = 2.4\text{ percent}.\]

v

Now add the variable prppov to the regression in part(iv). What happens to \(\hat{\beta}_{\text{prpblk}}\)?
The result of the new regression is:
\[
\hat{psoda} = -0.51 + 0.08prpblk + 0.14\log(income) + 0.40\text{prppov}
\]
\[N = 401 \quad R - squared = 0.08.\]

The coefficient on prpblk goes from 0.12 to 0.08 when prppov (the proportion of individuals living in poverty in a zip code) is included in the regression — the coefficient decreases.

vi

Find the correlation between \log(income) and prppov. Is it roughly what you expected?
Given the results of the above regressions (and intuition), I expected a negative correlation. I do not have enough intuition to expect a particular magnitude of the correlation (although I was not surprised to see a large correlation between the two). When I found the correlation using Stata’s correlate function, I get: -0.84.

vii

Evaluate the following statement: “Because \log(income) and prppov are so highly correlated, they have no business being in the same regression.”
My concise evaluation: this statement is nonsense. They can certainly be in the same regression. Their simultaneous inclusion, however, makes it more difficult to identify the independent effects of each on the price of soda. This does NOT mean that it is illegitimate to include both in a regression.