ECO375 Tutorial 4
Wooldridge: Chapter 6 and 7

Matt Tudball

University of Toronto St. George

October 6, 2017

ECO375 Tutorial 4

Welcome back!

Today's coverage:

- Chapter 6, #3 (in slides)
- Chapter 6, #8 (in slides)
- Chapter 6, C10 (in slides)
- Chapter 7, #4 (in slides)
- Chapter 7, #8 (in slides)

Chapter 6, #3

Using the data in RDCHEM.dta, the following equation was obtained by OLS:

$$\widehat{rdintens} = 2.613 + .0003\ sales - .000000007\ sales^2$$
$$(.429) \quad (.00014) \quad (.0000000037)$$
$$n = 32,\ R^2 = .1484$$

i) At what point does the marginal effect of *sales* on *rdintens* become negative?

Chapter 6, #3

Using the data in RDCHEM.dta, the following equation was obtained by OLS:

$$\widehat{rdintens} = 2.613 + .0003\ sales - .000000007\ sales^2$$
$$(.429) \quad (.00014) \quad (.0000000037)$$
$$n = 32,\ R^2 = .1484$$

i) At what point does the marginal effect of *sales* on *rdintens* become negative?

We can take the derivative of $\widehat{rdintens}$ with respect to *sales* and set it equal to 0: $.000000014\ sales^* = .0003$. Then we know that the point at which the marginal effect of *sales* becomes negative is $sales^* = 21,428.57$.

Chapter 6, #3

ii) Would you keep the quadratic term in the model? Explain.

Chapter 6, #3

ii) Would you keep the quadratic term in the model? Explain.

Probably. The t-statistic on $\hat{\beta}_{sales^2}$ is -.000000007/.0000000037 = -1.89, which is significant against the one-sided alternative $H_1 : \beta_{sales^2} < 0$.

Chapter 6, #3

iii) Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report the standard errors and the R^2.

Chapter 6, #3

iii) Define *salesbil* as sales measured in billions of dollars: $salesbil = sales/1000$. Rewrite the estimated equation with *salesbil* and $salesbil^2$ as the independent variables. Be sure to report the standard errors and the R^2.

$$\begin{aligned} \widehat{rdintens} &= 2.613 + .0003\ sales - .000000007\ sales^2 \\ &= 2.613 + .0003\ (1000 * salesbil) - .000000007(1000 * salesbil)^2 \\ &= 2.613 + .3\ salesbil - .007\ salesbil^2 \\ &\quad (.429) \qquad (.14) \qquad (.0037) \end{aligned}$$

Recall that $se(\hat{\beta}_j) = \hat{\sigma}/[SST_j(1-R^2)]^{1/2}$ **(3.58). Rescaling** *sales* **will have no effect on** $\hat{\sigma}$ **or** R^2 **since it does not change the fit of the regression. It will, however, affect** SST_{sales} **and** SST_{sales^2}**. Specifically,** $SST_{salesbil} = \sum_{i=1}^{n}(salesbil - \overline{salesbil})^2 = \sum_{i=1}^{n}(sales - \overline{sales})^2/1000^2 = SST_{sales}/1000^2$**. Similarly** $SST_{salesbil^2} = SST_{sales^2}/1000^4$**. Therefore we need to scale the standard errors of** $\hat{\beta}_{salesbil}$ **and** $\hat{\beta}_{salesbil^2}$ ***up*** **by** 1000 **and** 1000^2 **respectively.**

Chapter 6, #3

iv) For the purpose of reporting the results, which equation do you prefer?

Chapter 6, #3

iv) For the purpose of reporting the results, which equation do you prefer?

The equation in part (iii) is easier to read because it contains fewer zeros to the right of the decimal. Of course the interpretation of the two equations is identical once the different scales are accounted for.

Chapter 6, #8

Suppose we want to estimate the effects of alcohol consumption (*alcohol*) on college grade point average (*colGPA*). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called *attend*). A standardised test score (say, *SAT*) and high school GPA (*hsGPA*) are also available.

Chapter 6, #8

i) Should we include *attend* along with *alcohol* as explanatory variables in a multiple regression model? (Think about how you would interpret $\beta_{alcohol}$).

Chapter 6, #8

i) Should we include *attend* along with *alcohol* as explanatory variables in a multiple regression model? (Think about how you would interpret $\beta_{alcohol}$).

This is going to be a judgement call. We need to think about the causal pathways going from alcohol consumption to college GPA. It is very possible that alcohol consumption *alcohol* reduces lecture attendance *attend* which then reduces college GPA *colGPA*. Therefore if we control for *attend* in our regression we need to interpret $\beta_{alcohol}$ as the effect of alcohol consumption on college GPA other than the effects coming through lecture attendance. Our decision to include *attend* or not depends on whether we consider this an acceptable interpretation. If not, we may want to omit *attend*.

Chapter 6, #8

ii) Should *SAT* and *hsGPA* be included as explanatory variables? Explain.

Chapter 6, #8

ii) Should *SAT* and *hsGPA* be included as explanatory variables? Explain.

These are probably okay to include as explanatory variables since they are likely to be determined before alcohol consumption, meaning that we are not controlling for a potential causal pathway. *SAT* and *hsGPA* may relate to alcohol consumption (ex. students with lower SAT scores may be less interested in academics and more interested in partying) and they are definitely related to college GPA. Therefore we want to include them both as controls.

A potential causal pathway we are shutting down, however, is the long-term effects of alcohol consumption. Students who began drinking heavily in high school may have impaired their academic performance, lowering *SAT* and *hsGPA*, which then continues to impair their college GPA *colGPA*.

Chapter 6, C10

Use the data in BWGHT2.dta for this exercise.

i) Estimate the equation

$$\log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

by OLS, and report the results in the usual way. Is the quadratic term significant?

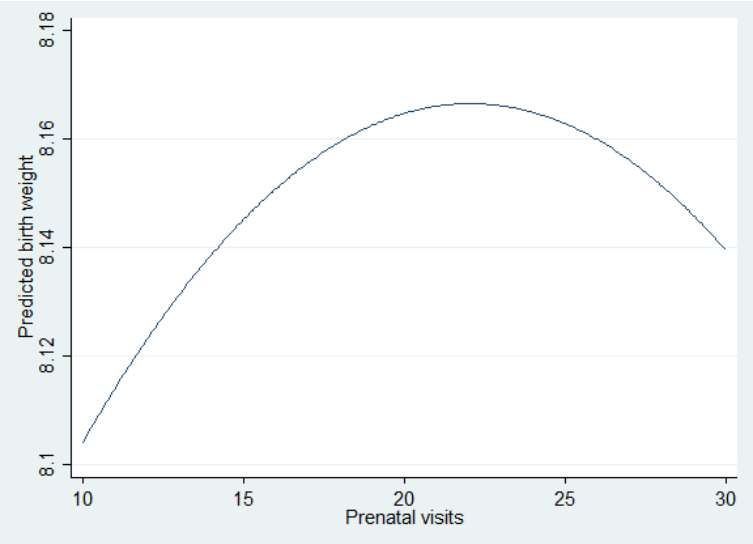
The results are displayed here:

$$\widehat{\log(bwght)} = \underset{(.0273)}{7.958} + \underset{(.00368)}{.0189}\ npvis - \underset{(.00012)}{.000429}\ npvis^2$$
$$n = 1764,\ R^2 = .0213$$

We can see that the t-statistic on $\hat{\beta}_2$ is -.000429/.00012 = -3.575, indicating that the quadratic term is very significant. Stata also reports a p-value of 0.000 (meaning it smaller than 0.001).

Chapter 6, C10

ii) Show that, based on the equation from part (i), the number of prenatal visits that maximises log(*bwght*) is estimated to be about 22. How many women had at least 22 prenatal visits in the sample?



Just like with Chapter 6 #3, we can take the derivative of the equation in (i) with respect to *npvis* and set it equal to 0. We know that this will be a maximum since the coefficient on $npvis^2$ is negative.

$.0189 - .000858\ npvis^* = 0$ **which indicates** $npvis^* = 22.02 \approx 22$**.**

Chapter 6, C10

iii) Does it make sense that birth weight is actually predicted to decline after 22 prenatal visits?

iii) Does it make sense that birth weight is actually predicted to decline after 22 prenatal visits?

While prenatal visits are a good thing for helping to prevent low birth weight, a woman's having many prenatal visits is a possible indicator of a pregnancy with difficulties. So it does make sense that the quadratic has a hump shape, provided we do not interpret the turnaround as implying that too many visits actually causes low birth weight.

iv) Add mother's age into the equation, using a quadratic functional form. Holding *npvis* fixed, at what mother's age is the birth weight of the child maximised? What fraction of women in the sample are older than the "optimal" age?

mage **is the variable indicating mother's age. We estimate that** $\hat{\beta}_{mage} = .0254$ **and** $\hat{\beta}_{mage^2} = -.000412$**. Similar to (i) we know that** *mage* **is maximised when** $.0254 - .000824\ mage^* = 0$**, which indicates** $mage^* = 30.83 \approx 31$**.**
By tabulating our data we can see that 66.98% of women in the sample are younger than 31, so 33.02% are older than 31.

Chapter 6, C10

v) Would you say that mother's age and number of prenatal visits explain a lot of the variation in log(*bwght*)?

No. $R^2 = .0256$ **so we are explaining only about 2.6% of the variation in log(*bwght*).**

Chapter 6, C10

vi) Using quadratics for both *npvis* and *mage*, decide whether using the natural log or the level of *bwght* is better for predicting *bwght*.

When we use *bwght* as a dependent variable instead of $log(bwght)$ we obtain a $R^2 = 0.0192$. However, to compare this to the R^2 coming from the regression with $log(bwght)$ as a dependent variable, we need to know how well that regression predicts *bwght* in levels (see section 6.4). We know that this is $\widehat{bwght} = exp(\hat{\sigma}^2/2)exp(\widehat{log(bwght)})$ (6.42). From here we want to compute the square correlation between $\widehat{bwght}$ and *bwght* (this is another way of calculating the R^2). I compute the correlation to be .1362 and so the square correlation is .0186.

This means that the regression with *bwght* as the dependent variable explains a tiny bit more of the variation (.0192) than the regression with $log(bwght)$ as a dependent variable (.0186).

Chapter 7, #4

An equation explaining chief executive officer salary is

$$\widehat{\log(salary)} = \underset{(.30)}{4.59} + \underset{(.032)}{.257}\ \log(sales) + \underset{(.004)}{.011}\ roe + \underset{(.089)}{.158}\ finance + \underset{(.085)}{.181}\ consprod - \underset{(.099)}{.283}\ utility$$
$$n = 209,\ R^2 = 0.357$$

The data used are in CEOSAL1.dta, where *finance*, *consprod* and *utility* are binary variables indicating the financial, consumer products and utilities industries. The omitted variable is transportation.

Chapter 7, #4

i) Compute the approximate percentage difference in estimated salary between the utility and transportation industries, holding *sales* and *roe* fixed. Is the difference statistically significant at the 1% level?

We can see from the estimated regression in the previous slide that the coefficient on *utility* is -.283. This means that CEO salaries in the utility industry are approximately 28.3% less on average than the CEO salaries in the transportation industry (which was omitted). The t-statistic on this coefficient is $-.283/.099 = -2.86$, which is very statistically significant.

Chapter 7, #4

ii) Use equation (7.10) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (i).

Recall that the exact percentage difference in salaries is $100 * [exp(\hat{\beta}_{utility}) - 1]$ **(7.10). See Example 7.5 for a derivation. The exact percentage difference between the utility and transportation industries is therefore -24.7% and so this estimate is somewhat smaller in magnitude than the one in (i).**

Chapter 7, #4

iii) What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.

The proportionate difference is .181 - .158 = .023, or about 2.3%. We could write a slightly different multiple regression equation,

$$\begin{aligned}\log(salary) =& \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \delta_1 consprod + \delta_2 utility \\ &+ \delta_3 trans + u\end{aligned}$$

Now *finance* is the omitted industry, so we would interpret δ_1 as the approximate percentage difference in CEO salaries between the consumer products and finance industries. We can tell if this difference is statistically significant by checking whether δ_1 is statistically significant.

Chapter 7, #8

Suppose you collect data from a survey on wages, education, experience and gender. In addition, you ask for information about marijuana useage. The original is: "On how many separate occasions last month did you smoke marijuana?"

i) Write an equation that would allow you to estimate the effects of marijuana useage on wages, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by $x\%$".

Chapter 7, #8

Suppose you collect data from a survey on wages, education, experience and gender. In addition, you ask for information about marijuana useage. The original is: "On how many separate occasions last month did you smoke marijuana?"

i) Write an equation that would allow you to estimate the effects of marijuana useage on wages, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by $x\%$".

$$\log(wage) = \beta_0 + \beta_1 useage + \beta_2 educ + \beta_3 exper + \beta_4 female + u$$

Chapter 7, #8

ii) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug useage for men and women?

Chapter 7, #8

ii) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug useage for men and women?

$$\begin{aligned}\log(wage) =& \beta_0 + \beta_1 useage + \beta_2 educ + \beta_3 exper + \beta_4 female \\ &+ \beta_5 useage \cdot female + u\end{aligned}$$

Testing that there are no difference in the effects of drug useage for men and women would involve testing $H_0 : \beta_5 = 0$ against $H_1 : \beta_5 \neq 0$.

Chapter 7, #8

iii) Suppose you think it is better to measure marijuana useage by putting people into one of four categories: non-user, light user (1 to 5 times per month), moderate user (6 to 10 times per month) and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana useage on wage.

Chapter 7, #8

iii) Suppose you think it is better to measure marijuana useage by putting people into one of four categories: non-user, light user (1 to 5 times per month), moderate user (6 to 10 times per month) and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana useage on wage.

Assuming no interaction effect between useage and sex, the model would look like,

$$\begin{aligned}\log(wage) =& \beta_0 + \beta_1 light + \beta_2 moderate + \beta_3 heavy + \beta_4 educ + \\ &+ \beta_5 exper + \beta_6 female + u\end{aligned}$$

In this model, non-user is the omitted category.

Chapter 7, #8

iv) Using the model in part (iii)m explain in detail how to test the null hypothesis that marijuana useage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

Chapter 7, #8

iv) Using the model in part (iii)m explain in detail how to test the null hypothesis that marijuana useage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

The null hypothesis here is $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. Naturally this is going to be an F-test on $q = 3$ restrictions. We are also going to have degrees of freedom $df = n - 6 - 1$ for a sample of size n, since we have 6 independent variables in the unrestricted model. So we would be obtaining a critical value from the $F_{q,n-7}$ distribution.

Chapter 7, #8

v) What are some of the potential problems with drawing causal inference using the survey data you collected?

Chapter 7, #8

v) What are some of the potential problems with drawing causal inference using the survey data you collected?

We can think of several here.

1. **Respondents may not accurately report their marijuana useage, perhaps out of social stigma or fear of legal repercussions.**
2. **There may be omitted variables which determine both marijuana useage and wages. For example, people living in urban areas may have easier access to marijuana and may earn higher wages on average. In this example, our estimate would be downward biased.**