Problem set 2: understanding ordinary least squares regressions

September 12, 2013

1 Introduction

This problem set is meant to accompany the undergraduate econometrics video series on youtube; covering roughly the 30th video through to the 85th.

2 NBA Wages - practical

In this part we are going to investigate the factors which affect NBA players' wages, as a means to helping us understand multiple regression analysis. The dataset we are going to use is prebuilt into Gretl, and comes from Jeffrey Wooldridge's great textbook, 'Introductory Econometrics'. In order to load the dataset, navigate to File → Open data → Sample file... Then select the Wooldridge tab (if you can see it - if you cannot I will explain how to access this in a minute) and open the 'nbasal' dataset. If you cannot see the Wooldridge tab you may need to go to:

`http://gretl.sourceforge.net/gretl_data.html#jw`

and download the wooldridge_data.exe or wooldridge.tar.gz file (I recommend the former method because it handles the installation for you). Alternatively, you can load in the sample dataset from clicking on the 'Look on server' button at the top of the sample datasets box. Alternatively, I have provided a link to download a .xls file containing the dataset on: (however note that this data is missing the descriptive labels of the native Wooldridge Gretl file)

`http://www.oxbridge-tutor.co.uk/#!datasets/culy`

The data contains demographic and position information for a sample of 269 players. The descriptive labels which are next to the variables contain descriptions of the series, and are a good idea for when you go on to do your own projects as they help you to keep track of variables.

1. Draw a boxplot for the players' wages. (If you don't know how to do a given plot etc. then consult the user manual by clicking the 'Help' menu). Which way are the players' wages skewed? Towards infinity or zero?

2. Let's investigate the relationships between variables in our dataset. In practice if two variables are highly correlated with one another, then we may run into the problems caused by multicollinearity. This will make it hard for ordinary least squares to decipher the effect of one variable from another in a regression model. One way of investigating the relationships is via their correlation. If you select the option of 'Correlation matrix' from the 'View' menu, and include all the variables in the dataset, this will output the bivariate correlations between all variables in the NBA dataset. This can be a useful tool to allow one to get a quick handle on the data by seeing how strong the relationship is between different variables in your dataset. Are there any variables that are particularly highly correlated with experience? What would be the issue of including both of these measures in a regression with wages as the dependent variable?

3. Another useful aspect of a correlation matrix is that it can give you a feel for which variables are correlated with your dependent variable. Which variables (other than the log of wages), show the highest correlation with wages?

4. Graphically investigate whether players who are more experienced earn more. How strong is the correlation between these two variables?

5. Create an ordinary least squares model which investigates how experience affects a player's wages. (Tip: You can save models for future use/viewing by clicking 'save as icon and close' in the model window. To access your model go to View → Icon view then click on the model. If you right click on a model you can change its name.)

6. What is the average wage increase associated with an increase in experience by one year implied by your model?

7. Do you think that the estimates of the effect of experience on wages is likely too big or too small? Which Gauss-Markov assumption is being violated, and why?

8. Create another regression with wages as a dependent variable, and age as the independent variable (along with a constant). Does this imply that the effect of age is positive or negative?

9. What would would be the average wage implied by your model for an individual of 30? What about for a 90 year old? What is the problem with the latter estimate?

10. How might we rectify the issue of the unrealistic estimates from the previous model?

11. Now create a regression with *both* experience and age in the model. What has happened to the sign of the coefficient on age? Why has this happened?

12. Let's now try to examine whether individuals who score more points tend to earn more by creating a regression of wages on points per game (and a constant). You can view a graph of the fitted regression line by navigating to Graphs → Fitted, actual plot from within the model window. The option you should select is 'against points'. From this you can see a graph of actual vs predicted wages vs points. What does your model suggest would be the increase in wages for an increase in 10 points per game?

13. You can look at a graph of the residuals (the estimated errors) from the regression, by clicking into the model (if you are not already in the model window), and navigating to Graphs → Residual plot. There are then a number of different options available, from seeing a plot of residuals against observation number to seeing a plot against the values of experience. Which of these plots should you use to graphically inspect for heteroscedasticity?

14. Does there appear to be heteroscedasticity? What might be causing it? How might we rectify it?

15. Do you think that the effect of points on wages predicted by your model is too high or low? Why might this be the case?

16. We are now going to create two new variables in Gretl: 'pointsq' and 'pointsc' equal to the square of points and its cube respectively. To do this navigate to Add → Define new variable... This allows a user to enter a formula for the construction of a new variable from an old one. For example to create 'pointsq', you can enter the formula: 'pointsq = points^2'. Here the '^' means 'raise that variable to the power'. Go ahead and create 'pointsq' and 'pointsc'.

17. Now create two new regression models (keeping your current regression of wages on points):

$$wage_i = \alpha + \beta_1 points_i + \beta_2 pointsq_i$$

$$wage_i = \alpha + \beta_1 points_i + \beta_2 pointsq_i + \beta_3 pointsc_i$$

    (a) Which of these regressions has the highest value of R-squared? What does this mean?
    (b) What is the interpretation of the coefficient on 'pointsc' in the last regression model?
    (c) Which of these regressions has the highest value of adjusted R-squared?
    (d) Out of the three specifications, which would you prefer?

18. Freestyle: try to create a model which you believe captures explains players wages best in terms of other attributes

3 Theory

1. A researcher is interested in quantifying the effect of the number of broken windows in a block on property prices, and the results of a preliminary regression are:

   $$Hprice_i = 100 - 10windows_i$$

   Where $windows_i$ represents the number of broken windows counted on a block, i, and $Hprice_i$ is the average property value (in thousands of \$) on that same block.

   (a) What is the interpretation of the coefficient on $windows_i$?

   (b) What are the causes of endogeneity in this model?

   (c) Do you think that this coefficient over or under estimates the effect of broken windows on house prices?

   (d) Another variable is included in the regression, $emerg_i$, which is a measure of the number of emergency services calls which were made from each block over a period of time. And the result of the regression is:

   $$Hprice_i = 100 - 3windows_i - 5emerg_i$$

   Why has the coefficient on $windows_i$ fallen relative to what it was before in this new model? Not only this, but neither coefficients on $windows_i$ or $emerg_i$ are statistically significant. However, both coefficients jointly appear to be significant in determining house prices. Why might this be?

2. The zero conditional mean assumption of the Gauss-Markov conditions is often stated as:

   $$\mathbb{E}[\varepsilon_i|X_i] = 0 \qquad (1)$$

   (a) Can you prove that this implies that $\mathbb{E}[\varepsilon_i X_i] = 0$?

   (b) Does this imply that X_i and ε_i are uncorrelated? Prove it.

   (c) Does the covariance between X_i and ε_i being zero imply independence of these variables?

3. A researcher is interested in measuring what the effect of an individual's innate 'language intelligence' is on their ability to learn a language. She finds 100 volunteers for the study who have all not learned French before, nor have they learned any other languages to any serious fluency. Her theory is that those individuals who have higher innate measures of 'language intelligence' will take less time to reach of level of proficiency in French.

   Each volunteer is enrolled in a day course in basic French, and is tested at the end of the day in their progress in the language. At the end of the day each participant also takes a standardised IQ test. The researcher then carries out the following regression:

   $$score_i = \alpha + \beta IQ_i + u_i$$

   (a) Do you think that β fairly represents the effect of an incremental point of IQ on an individual's performance in the end of day test? Why/why not?

(b) The above equation is amended to include any other relevant explanatory variables. The researcher is aware that IQ is not a perfect measure of an individual's 'language intelligence'. However, she supposes that it is an adequate proxy - meaning that it is not a biased estimate of 'language intelligence'. Will the least squares estimator $\hat{\beta}$ be unbiased?

(c) Prove either way your answer for the last question.