Introductory Econometrics

Tutorial 5

<u>PART A:</u> To be done before you attend the tutorial. The questions are similar to questions on the mid-semester test.

In all three of the following questions, the correct answer (e). However, the goal is not to spot the correct answer, but to understand why other answers are wrong. Write a sentence or two on each suggested answer that explains why that answer is correct or incorrect.

1. We have estimated the model $wage = \beta_0 + \beta_1$ experience $+u$ using OLS based on a sample of 4 observations. We know that the matrix of explanatory variables is

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$$

and we are told that the OLS residuals are:

$$\hat{\mathbf{u}} = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$$

We can immediately say that the residual vector:

(a) is reported correctly because it sums to zero

(b) is reported incorrectly because it should be a 2×1 vector

(c) is reported correctly but shows a poor fit because $\mathbf{X}'\hat{\mathbf{u}} \neq \mathbf{0}$

(d) is reported correctly because it is linearly independent of columns of $\mathbf{X}$

(e) is reported incorrectly because it is not orthogonal to the second column of $\mathbf{X}$

2. The multiple regression model in matrix form is

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}}\underset{(k+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{u}}$$

where dimensions are specified below each vector and matrix. We denote the estimated model by

$$\underset{n\times 1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}}\underset{(k+1)\times 1}{\widehat{\boldsymbol{\beta}}} + \underset{n\times 1}{\widehat{\mathbf{u}}}$$

in which $\widehat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta}$ and $\widehat{\mathbf{u}}$ is the vector of OLS residuals. The vector of predicted values given by the OLS, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ is the linear combination of columns of $\mathbf{X}$ that is closest to $\mathbf{y}$. This implies that:

(a) $\mathbf{X}'\widehat{\mathbf{y}} = \mathbf{0}$

(b) $\mathbf{X}'\mathbf{y} = \mathbf{0}$

(c) $\mathbf{X}'\mathbf{u} = \mathbf{0}$

(d) $E(\widehat{\mathbf{y}} \mid \mathbf{X}) = \mathbf{0}$

(e) $\mathbf{X}'\widehat{\mathbf{u}} = \mathbf{0}$

3. In the multiple regression model shown in the previous question, which one of the following statements is **incorrect**:

   (a) $\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$

   (b) The sum of squared residuals is the square of the length of the vector $\widehat{\mathbf{u}}$

   (c) The residual vector is orthogonal to each of the columns of $\mathbf{X}$

   (d) The square of the length of $\mathbf{y}$ is equal to the square of the length of $\widehat{\mathbf{y}}$ plus the square of the length of $\widehat{\mathbf{u}}$ by the Pythagoras theorem

   (e) $\frac{1}{n}\sum_{i=1}^{n}u_i = 0$

Do not forget to bring your answers to PART A and a copy of the tutorial questions to your tutorial.

<u>Part B:</u> This part will be covered in the tutorial. It is still a good idea to attempt these questions before the tutorial.

1. Consider the problem of measuring the comovement of a particular stock's excess return (Qantas) with the excess return of a market portfolio (AllOrds). Excess return is the return minus risk free rate of return (say 3 month term deposit rate). We denote the stock's excess return by y and the market portfolio's excess return by x. We use data on the previous n months on y and x and we use the linear model:

$$y_t = \beta_0 + \beta_1 x_t + u_t \quad t = 1, \ldots, n$$

   which in matrix notation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{1}$$

   We use the index t because we are using time series observations. We assume that $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$, which means that the part of the movement in Qantas shares not related to current month's market events are completely related to events specific to Qantas, i.e. it is also not predictable using previous months or future months market events. Consider the following estimator of the slope parameter β_1 (in our sample, $x_n \neq x_1$)

$$\begin{aligned}\tilde{\beta}_1 &= \frac{y_n - y_1}{x_n - x_1}, \text{ the slope of the line connecting the first observation to the last observation} \\ &= \frac{\triangle \text{ in Qantas excess return over the } n \text{ months in this sample}}{\triangle \text{ in AllOrds excess return over the } n \text{ months in this sample}}\end{aligned}$$

   (a) Derive the expected value of $\tilde{\beta}_1$ conditional on $\mathbf{X}$ to show that this is an unbiased estimator of β_1.

   (b) Assuming $Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2\mathbf{I}_n$, can $\tilde{\beta}_1$ have a smaller variance than the OLS estimator? Explain.

2. Problem C2, Chapter 3 of the textbook: Use the data in HPRICE1.WF1 to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

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

   i) Write out the results in equation form.

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

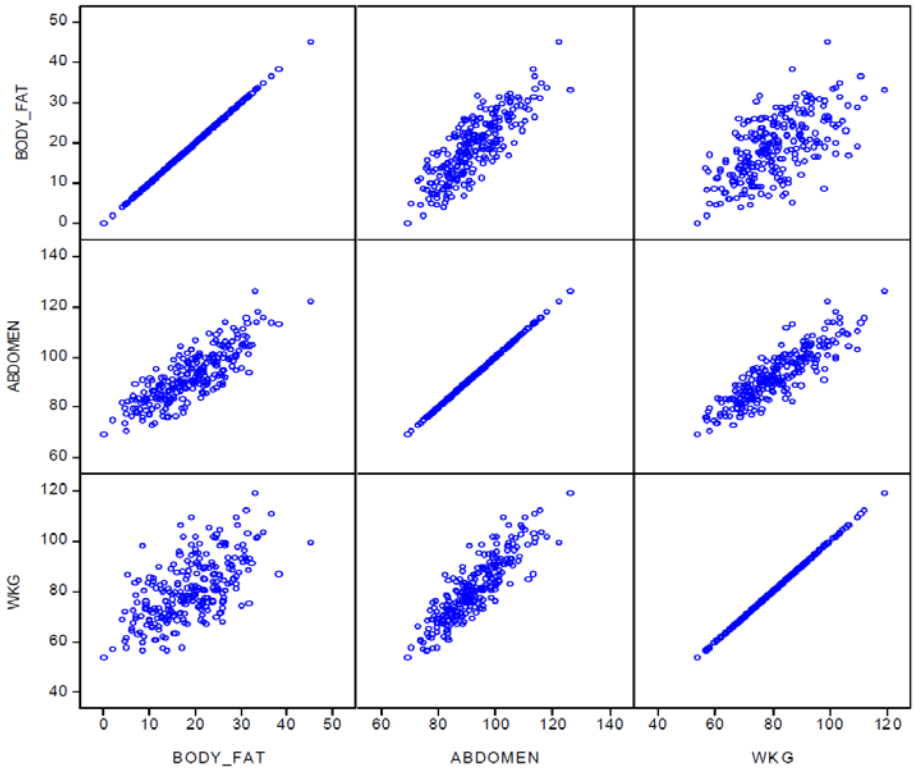
iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

v) The first house in the sample has sqrft = 2,438 and bdrms = 4. Find the predicted selling price for this house from the OLS regression line.

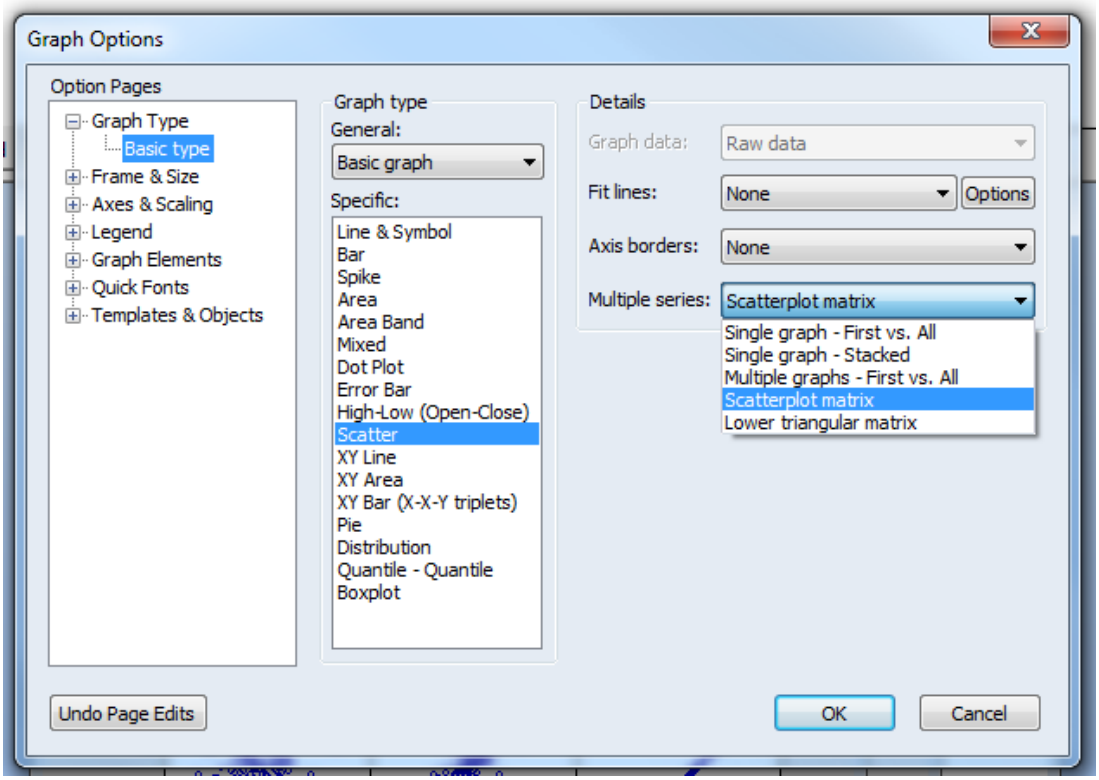
vi) The actual selling price of the first house in the sample was \$300,000 (so price = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

3. We would like to make an "app" where users input their easy to measure body characteristics and the app predicts their body fat percentage. We start with making an app for men. We have data on body fat percentage (BODY_FAT), weight in kg (WKG) and abdomen circumference in cm (ABDOMEN) for 251 adult men. The matrix of scatter plots of each pair of these three variables in our sample is given below.



The plots in the first row are: the scatter plot of body fat against body fat (which is the 45 degree line) at the left corner, the scatter plot of body fat against abdomen circumference in the middle, and the scatter plot of body fat against weight in the top right corner. You can create these matrices in Eviews by graphing more than two variables and then choosing scatter plot,

with the scatter plot matrix option, as shown in the screen shot below.



Without estimating any regressions, explain what these plots can tells us about each of the following (the correct answer for one of these is "nothing"):

(a) the sign of the coefficient of ABDOMEN in a regression of BODY_FAT on a constant and ABDOMEN,

(b) the sign of the coefficient of WKG in a regression of BODY_FAT on a constant and WKG,

(c) which of the two regressions explained in parts (a) and (b) is likely to have a better fit,

(d) the sign of the coefficient of WKG in a regression of BODY_FAT on a constant, ABDOMEN and WKG.

4. With the same data as above, we have estimated three regressions:

$$\begin{aligned}
\widehat{BODY_FAT} &= -12.63 + 0.39WKG, \qquad R^2 = 0.385,\ \bar{R}^2 = 0.382 \\
\widehat{BODY_FAT} &= -38.60 + 0.62ABDOMEN, \quad R^2 = 0.681,\ \bar{R}^2 = 0.679 \\
\widehat{BODY_FAT} &= -42.94 + 0.91ABDOMEN - 0.27WKG,\ R^2 = 0.724,\ \bar{R}^2 = 0.722
\end{aligned}$$

(a) The signs and the R^2s of the first two regressions must agree with your answers to parts (a), (b) and (c) of the previous question. If they don't, then discuss these in the tutorial or during consultation hours.

(b) Think about the negative coefficient of WKG in the third equation. Does it make sense? (Hint: yes, it makes very good sense, and it highlights the extra information that multiple regression extracts from the data that simple two variable regressions cannot do). Explain, to a non-specialist audience, what the estimated coefficient of WKG in the third regression tells us.

(c) If weight was measured in pounds rather than kilograms (each kilogram is 2.2 pounds), how would the above regression results change? Check your answers by running the regressions using bodyfat.wf1 file.