Introductory Econometrics

Tutorial 5

<u>PART A:</u> To be done before you attend the tutorial. This is very similar to the first question on the first assignment, and will be good preparation for the mid-semester test as well. Solutions will be provided after the first assignment is handed in. The question asks you to derive the expected value of four estimators conditional on explanatory variables. Since the derivation of these conditional expectations involve similar steps, you can divide the task between your teammates and each derive the conditional expectation for only one of the estimators, and hand that one to your tutor.

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. Four estimators for the slope parameter β_1 have been proposed (in our sample, $\bar{x} \neq 0$, $x_n \neq x_1$ and $x_n \neq \bar{x}$, so all of the proposed estimators are well defined):

1. $\hat{\beta}_1^{[1]} = \frac{\bar{y}}{\bar{x}} = \frac{\frac{1}{n}\sum_{t=1}^n y_t}{\frac{1}{n}\sum_{t=1}^n x_t}$, the ratio of the average excess return of Qantas shares to the average excess return of the market portfolio,

2. $\hat{\beta}_1^{[2]} = \frac{y_n - y_1}{x_n - x_1}$, the slope of the line connecting the first observation (x_1, y_1) to the last observation (x_n, y_n),

3. $\hat{\beta}_1^{[3]} = \frac{y_n - \bar{y}}{x_n - \bar{x}}$, the slope of the line connecting the last observation (x_n, y_n) to the sample average point $(\bar{x}, \bar{y})$,

4. $\hat{\beta}_1^{[4]} = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x^2}$, the ratio of the sample covariance between x and y to the sample variance of x, which is the OLS estimator of β_1.

Divide the above among the 4 members of your team. Each member should derive the expected value of a different one of the above estimators conditional on $\mathbf{X}$ to determine if that estimator is an unbiased estimator of or not. Conditioning on $\mathbf{X}$ allows you to treat $x_1, x_2, \ldots, x_n$ and any expression that only involves these, such as $\bar{x}$, as constants. Although each of you are welcome to attempt all four, but it is sufficient to submit your work on one of the above.

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

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

1. In the set up of Part A, consider a fifth estimator, which is the average slope of n lines connecting each of the n sample points (x_t, y_t) for $t = 1, \dots, n$ to the sample average point $(\bar{x}, \bar{y})$, assuming that $\bar{x}$ is not equal to any x_t for $t = 1, \dots, n$.

   - Assume a small n, say $n = 4$, and show these lines graphically (to convince yourself that this is not a crazy estimator), and also express this estimator algebraically.
   - Derive its expected value conditional on $\mathbf{X}$ to show that this is an unbiased estimator of β_1.
   - Assuming $Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$, can this estimator have a smaller variance than the OLS estimator (which was denoted by $\hat{\beta}_1^{[4]} = \frac{\hat{\sigma}_{x,y}}{\hat{\sigma}_x^2}$ in part A)? 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).

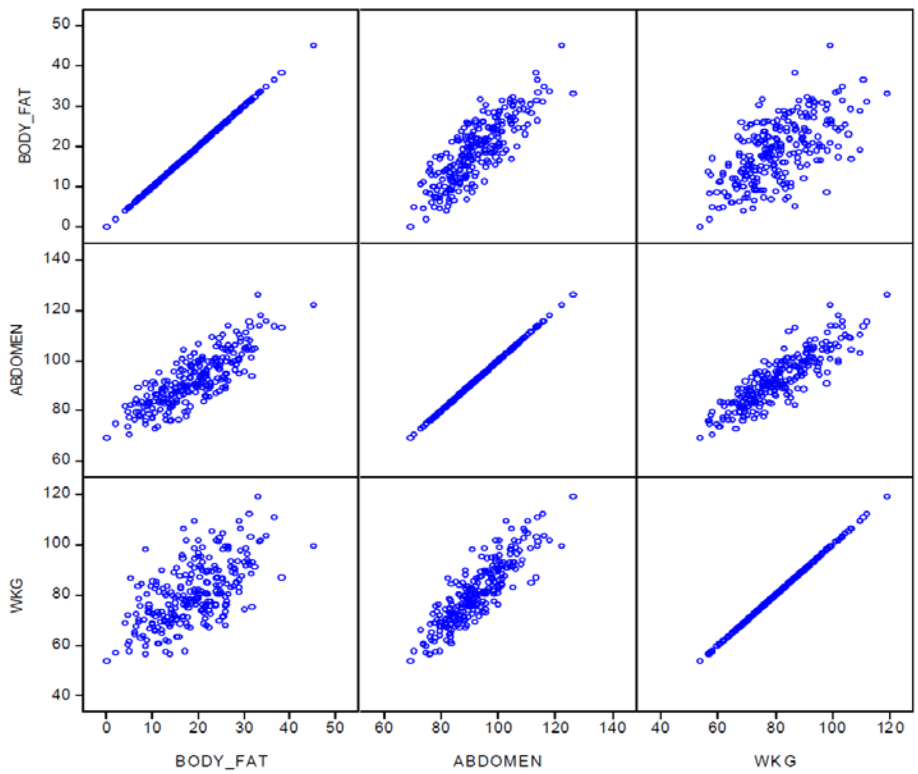
iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

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.

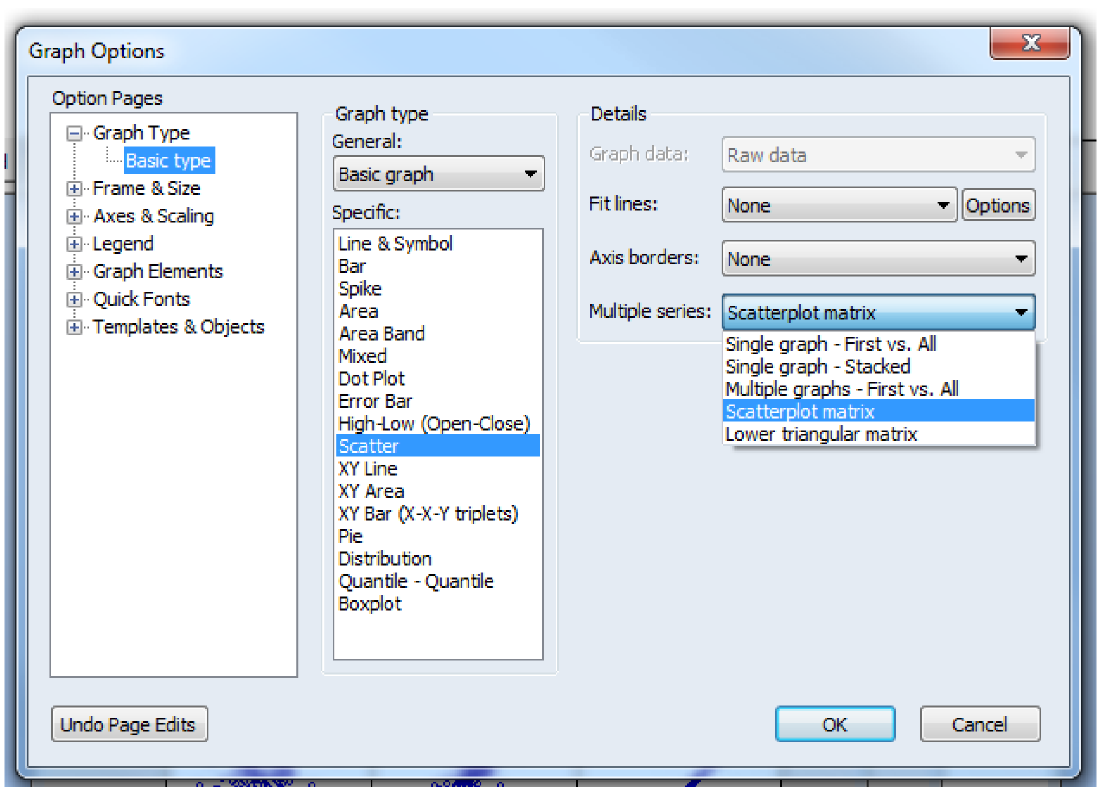
vi) The actual selling price of the first house in the sample was \$300,000 (so price = 300). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

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.

(d) If body fat was regressed on a constant only, what would the OLS estimate of the constant be? Answer it first and then check your answer using bodyfat.wf1 file.