**Introductory Econometrics**
Tutorial 5 Solutions

**PART A:**

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: Summing to zero only implies that $\hat{\mathbf{u}}$ is orthogonal to the first column of $\mathbf{X}$,

$$\begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} = 0$$

but the OLS residual vector has to be orthogonal to all columns of $\mathbf{X}$

(b) is reported incorrectly because it should be a $2\times1$ vector: This is rubbish. Every observation will have a residual, so the dimension of $\hat{\mathbf{u}}$ is the same as the dimension of $\mathbf{y}$

(c) is reported correctly but shows a poor fit because $\mathbf{X}'\hat{\mathbf{u}} \neq \mathbf{0}$ : This cannot be true because if $\hat{\mathbf{u}}$ was correct $\mathbf{X}'\hat{\mathbf{u}}$ would be equal to zero

(d) is reported correctly because it is linearly independent of columns of $\mathbf{X}$ : while it is true that $\hat{\mathbf{u}}$ is linearly independent of columns of $\mathbf{X}$, but any vector that is not a linear combination of columns of $\mathbf{X}$ would be linearly independent of columns of $\mathbf{X}$. But not all such vectors are orthogonal to columns of $\mathbf{X}$.

(e) is reported incorrectly because it is not orthogonal to the second column of $\mathbf{X}$ : yes, this is the correct answer because:

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

2. The multiple regression model in matrix form is

$$\underset{n\times1}{\mathbf{y}} = \underset{n\times(k+1)}{\mathbf{X}} \underset{(k+1)\times1}{\boldsymbol{\beta}} + \underset{n\times1}{\mathbf{u}}$$

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) $\mathbf{X}'\widehat{\mathbf{y}} = \mathbf{0}$ : this is rubbish. $\widehat{\mathbf{y}}$ is a linear combination of columns of $\mathbf{X}$, that is, it lies in the same plane as columns of $\mathbf{X}$. Therefore, it cannot be orthogonal to columns of $\mathbf{X}$

(b) $\mathbf{X}'\mathbf{y} = \mathbf{0}$ : this is rubbish also. $\mathbf{X}$ is the data matrix on independent variable, and $\mathbf{y}$ is the vector of observed dependent variables. There is no chance that these numbers will be such that $\mathbf{y}$ will be orthogonal to columns of $\mathbf{X}$.

(c) $\mathbf{X}'\mathbf{u} = \mathbf{0}$ : this is trickier. We know that $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$, which implies $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$. But if the expected value of a random variable is zero, it does not imply that that random variable is always zero. So, $\mathbf{X}'\mathbf{u}$ is a random variable (actually a $(k+1)\times 1$ vector of random variables) with mean $\mathbf{0}$, so $\mathbf{X}'\mathbf{u} = \mathbf{0}$ is incorrect.

(d) $E(\widehat{\mathbf{y}} \mid \mathbf{X}) = \mathbf{0}$ : this is incorrect. $E(\widehat{\mathbf{y}} \mid \mathbf{X}) = E\left(\mathbf{X}\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \mathbf{X}E\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \mathbf{X}\boldsymbol{\beta}$. This in general is not zero.

(e) $\mathbf{X}'\widehat{\mathbf{u}} = \mathbf{0}$ : this is correct, because the shortest distance from $\mathbf{y}$ to column space of $\mathbf{X}$ is the vector $\widehat{\mathbf{u}}$ that must be perpendicular to the columns of $\mathbf{X}$ (otherwise it won't be the shortest route).

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

(a) $\frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$ : this is correct because we have

$$\begin{pmatrix}1 & 1 & \cdots & 1 & 1\end{pmatrix}\begin{pmatrix}\widehat{u}_1\\ \widehat{u}_2\\ \vdots\\ \widehat{u}_{n-1}\\ \widehat{u}_n\end{pmatrix} = \sum_{i=1}^{n}\widehat{u}_i = 0 \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$

(b) The sum of squared residuals is the square of the length of the vector $\widehat{\mathbf{u}}$ : That is true because the square of the length of any vector is sum of squares of its individual components

(c) The residual vector is orthogonal to each of the columns of $\mathbf{X}$ : that is true by the construction of OLS.

(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: this is true because $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{u}}$ are orthogonal to each other and therefore $\mathbf{y}$, $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{u}}$ form a right angled triangle.

(e) $\frac{1}{n}\sum_{i=1}^{n}u_i = 0$ : this is not true. Again we know $E(u_i) = 0$ for $i = 1, \ldots n$, but that does not imply that the average of any sample of $n$ observation on $u_i$ will be zero.

**PART B:**

1. This exercise shows that there can be reasonable alternative unbiased estimators for the slope parameter. It is instructive to notice the steps on the proof of unbiasedness: First, in the formula for the estimator, substitute for $y$ using the population model and simplify such that to get the parameter of interest (here $\beta_1$) on its own. Then take expectation of both sides and use $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$ to show that all terms that are added to $\beta_1$ have expected value zero. If that is possible, then you have shown that the estimator is unbiased. If it is not possible, then the estimator will be biased.

   (a)

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{y_n - y_1}{x_n - x_1} = \frac{\beta_0 + \beta_1 x_n + u_n - \beta_0 - \beta_1 x_1 - u_1}{x_n - x_1} \\
&= \frac{\beta_1 (x_n - x_1) + u_n - u_1}{x_n - x_1} = \beta_1 + \frac{u_n - u_1}{x_n - x_1} \\
E\left(\tilde{\beta}_1\right) &= \beta_1 + E\left(\frac{u_n - u_1}{x_n - x_1}\right) = \beta_1 + E\left(\frac{u_n}{x_n - x_1}\right) - E\left(\frac{u_1}{x_n - x_1}\right) \\
E(\mathbf{u} \mid \mathbf{X}) &= \mathbf{0} \Rightarrow E\left(\frac{u_t}{x_n - x_1}\right) = 0 \text{ for all } t \\
&\Rightarrow E\left(\tilde{\beta}_1\right) = \beta_1, \text{ i.e. } \tilde{\beta}_1 \text{ is unbiased}
\end{aligned}$$

   (b) This estimator cannot have a smaller variance than the OLS estimator because under the assumptions given in the question, Gauss-Markov Theorem tells us that OLS has the smallest variance among all linear unbiased estimators.

2. 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, *sqrft* is the area of the house in square feet, and *bdrms* is the number of bedrooms.

i) Write out the results in equation form.

$$\begin{aligned}
\widehat{price} &= -19.32 + 0.128 sqrft + 15.20 bdrms \\
n &= 88, \quad R^2 = 0.632
\end{aligned}$$

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

$\$15,200$

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).

$\Delta\widehat{price} = 0.128\Delta sqrft + 15.20\Delta bdrms = 0.128(140) + 15.20 = 33.12$ or $\$33,120$.

In part (i) a bedroom was added by making other rooms smaller (since size was kept constant). Here, the size is also increasing, which adds more to the value of the house.

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

63.2%

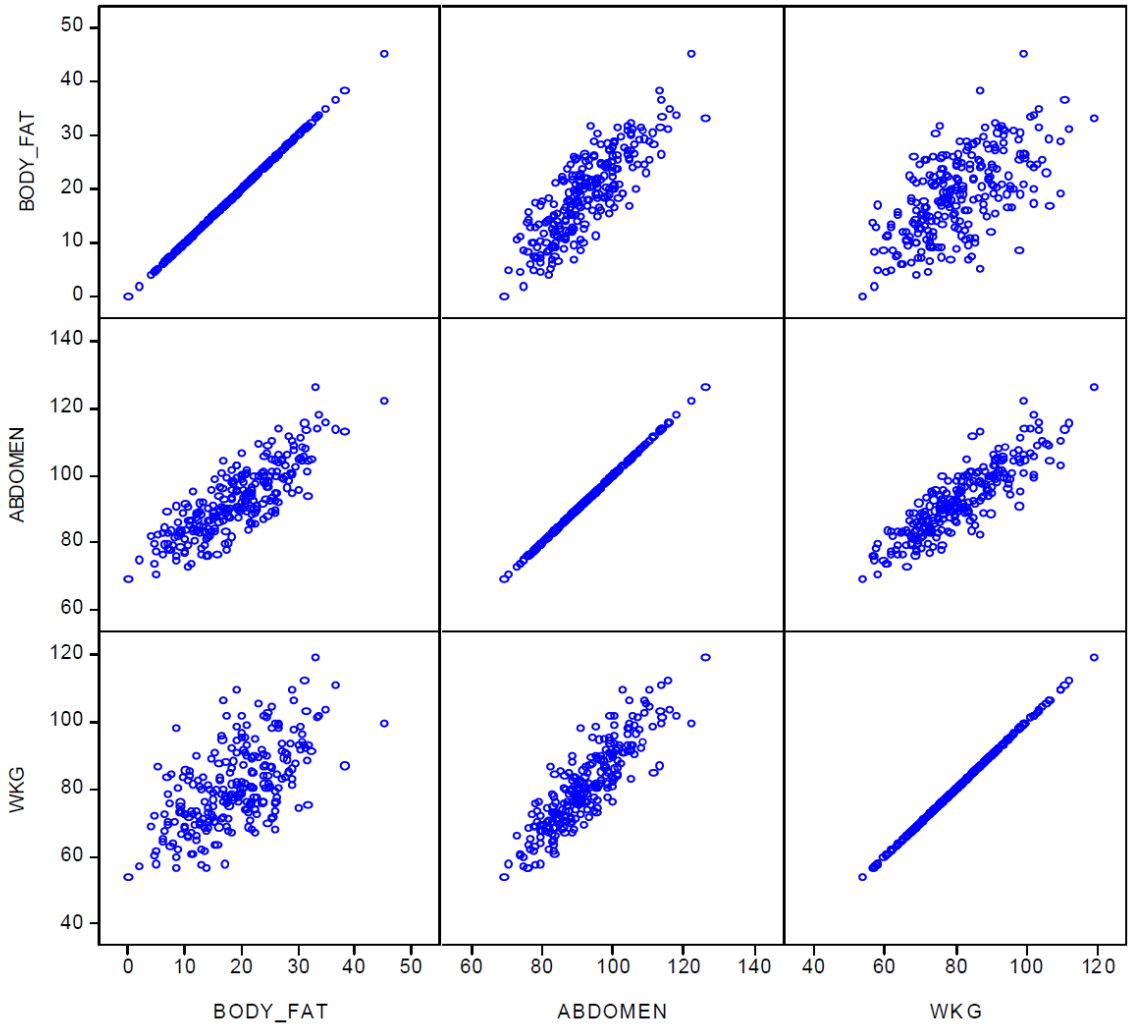
v) The first house in the sample has $sqrft = 2{,}438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

$$-19.32 + 0.128(2438) + 15.20(4) = 353.544, \text{ or } \$353,544.$$

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?

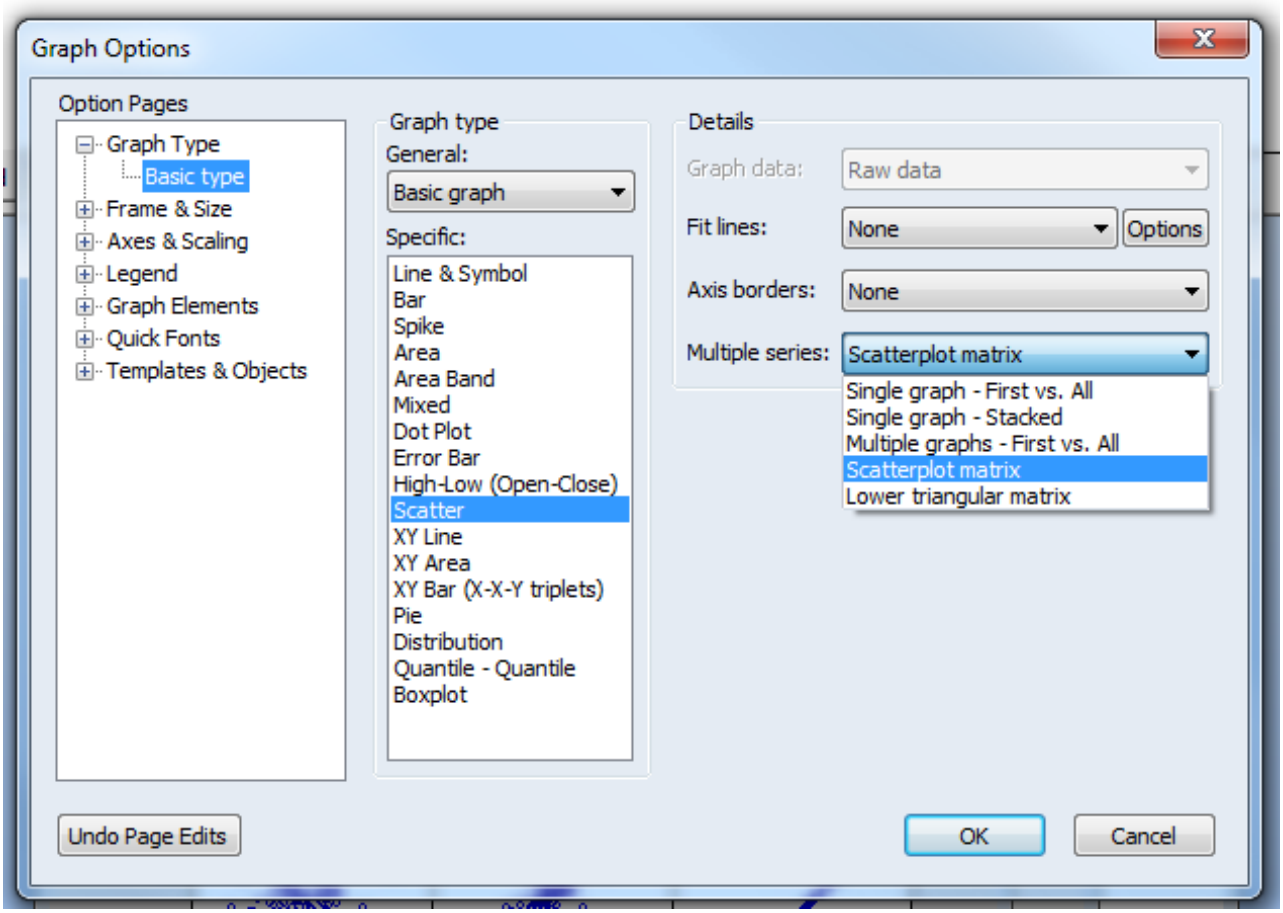
The buyer paid less than predicted. But there are many other features that we have not taken into account e.g. number of bathrooms, age of the house, whether it has been renovated or not, etc.

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,

$$\hat{\beta} = \frac{Cov(\widehat{ABDOMEN, BODY\_FAT})}{Var(\widehat{ABDOMEN})}$$

The scatter plot shows positive association, so sample covariance is positive therefore, the sign of $\hat{\beta}$ will be positive

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

$$\hat{\beta} = \frac{Cov(\widehat{WKG, BODY\_FAT})}{Var(\widehat{WKG})}$$

The scatter plot shows positive association, so sample covariance is positive therefore, the sign of $\hat{\beta}$ will be positive

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

In the scatter plot of body fat against abdomen, body fat values seem to be less dispersed around the mean for each value of abdomen circumference.
So, this regression 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.

Scatter plots cannot tell us anything about the correlation of body fat and weight after the influence of abdomen has been taken out.

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 $R^2$s 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.

They do :-)

(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.

If you think about it, it does! Two people with the same abdomen circumference, the one who is heavier is likely to be more athletic, (because muscle is heavier than fat) and therefore is likely to have less body fat.

(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.

The coefficient of $WKG$ in the first and the third equation will be divided by 2.2
All other estimated coefficients and the values of $R^2$ in all equation will stay the same

Dependent Variable: BODY_FAT
Method: Least Squares
Sample: 1 252 IF WEIGHT<300
Included observations: 251

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -42.94397 | 2.439845 | -17.60111 | 0.0000 |
| ABDOMEN | 0.905739 | 0.051864 | 17.46376 | 0.0000 |
| WKG | -0.269247 | 0.043113 | -6.245105 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.723970 | Mean dependent var | 18.87928 |
| Adjusted R-squared | 0.721744 | S.D. dependent var | 7.709026 |
| S.E. of regression | 4.066509 | Akaike info criterion | 5.655327 |
| Sum squared resid | 4101.050 | Schwarz criterion | 5.697464 |
| Log likelihood | -706.7435 | Hannan-Quinn criter. | 5.672284 |
| F-statistic | 325.2268 | Durbin-Watson stat | 1.790391 |
| Prob(F-statistic) | 0.000000 | | |

Dependent Variable: BODY_FAT
Method: Least Squares
Sample: 1 252 IF WEIGHT<300
Included observations: 251

| Variable | Coefficient | Std. Error | t-Statistic | Prob. |
|---|---|---|---|---|
| C | -42.94397 | 2.439845 | -17.60111 | 0.0000 |
| ABDOMEN | 0.905739 | 0.051864 | 17.46376 | 0.0000 |
| WKG*2.2 | -0.122385 | 0.019597 | -6.245105 | 0.0000 |

| | | | |
|---|---|---|---|
| R-squared | 0.723970 | Mean dependent var | 18.87928 |
| Adjusted R-squared | 0.721744 | S.D. dependent var | 7.709026 |
| S.E. of regression | 4.066509 | Akaike info criterion | 5.655327 |
| Sum squared resid | 4101.050 | Schwarz criterion | 5.697464 |
| Log likelihood | -706.7435 | Hannan-Quinn criter. | 5.672284 |
| F-statistic | 325.2268 | Durbin-Watson stat | 1.790391 |
| Prob(F-statistic) | 0.000000 | | |