## MiniQuiz - Week 5

1. **01 - Week 5**

MULTI | 1.0 point | 0.10 penalty | Single | Shuffle

Using data on weekly wage for a random sample of 400 full time employed individuals in Australia in 2019, we have estimated the following equation using OLS:

$$\widehat{wage_i} = 1856.79 - 356.94 female_i$$

where $female_i$ is a binary variable that is equal to 1 if the individual is female and is equal to 0 otherwise. These results reveal that:

(a) the average weekly wage of men in this sample is \$1856.79 and the average weekly wage of women is \$356.94.
(b) the average weekly wage of women in this sample is \$1856.79 and the average weekly wage of men is \$1499.85.
(c) every additional female employee costs \$356.94 per week less than the previous female employee.
(d) the average weekly wage of all individuals in the sample is \$1856.79, and the average weekly wage of all women in the sample is \$356.94 less than the overall average.
(e) the average weekly wage of men in this sample is \$1856.79 and the average weekly wage of women is \$1499.85. (100%)

*$\hat{\beta}_0$ is the sample average wage of men, and $\hat{\beta}_0 + \hat{\beta}_0$ is the sample average wage of women*

2. **02 - Week 5**

MULTI | 1.0 point | 0.10 penalty | Single | Shuffle

Using data on weekly wage for a random sample of 274 full-time employed male individuals in Australia in 2019, we have estimated the following equation using OLS:

$$\widehat{wage_i} = -1087.74 + 189.75 educ_i + 29.50 exper_i$$

where *educ* is years of education and *exper* is years of labour market experience. Which one of the following is the correct interpretation of the estimated coefficient of *educ*?

(a) An extra year of education increases wage of a full-time employed man by \$189.75.
(b) An extra year of education increases weekly wage of a full-time employed man by \$189.75.

(c) An extra year of education increases the predicted weekly wage of a full-time employed man by $189.75.
(d) In this sample, the average weekly wage of educated individuals is $189.75 higher than the average weekly wage of uneducated individuals.
(e) For two full-time employed men with the same years of experience, the predicted weekly wage of the person with one more year of education is $189.75 higher than the predicted weekly wage of the other person. (100%)

*Do not forget the "all else constant"*

3. **03 - Week 5**

MULTI | 1.0 point | 0.10 penalty | Single | Shuffle

Suppose that the R-squared associated with the estimated regression equation in the previous question is 0.45. Which of the following statements is correct?

(a) The explanatory variables explain 45% of the variation in the predicted weekly wage in the population.
(b) The explanatory variables explain 45% of the variation in the weekly wage in the population.
(c) The explanatory variables explain 45% of the variation in the weekly wage in the sample. (100%)
(d) The explanatory variables explain 45% of the variation in the predicted weekly wage in the sample.
(e) The explanatory variables explain 55% of the variation in the weekly wage in the sample.

4. **04 - Week 5**

MULTI | 1.0 point | 0.10 penalty | Single | Shuffle

Let $x$ be a random variable with mean $\mu$ and variance $\sigma^2$. Based on a random sample of 5 observations $\{x_1, x_2, x_3, x_4, x_5\}$ from the population of $x$, the following estimators for $\mu$ have been suggested:

$$\text{a) } \hat{\mu} = \frac{1}{5}\sum_{i=1}^{5} x_i, \text{ b) } \tilde{\mu} = \frac{1}{3}(x_1 + x_3 + x_5), \text{ and c)} \overline{\mu} = \frac{1}{2}(x_2 + x_4).$$

Which of the above are unbiased estimators for $\mu$?

(a) $\hat{\mu}$ only.
(b) None of them.
(c) $\tilde{\mu}$ only.
(d) $\overline{\mu}$ only.
(e) All three. (100%)

*$E(\hat{\mu}) = \frac{1}{5}\sum_{i=1}^{5} E(x_i) = \mu$, $E(\tilde{\mu}) = \frac{1}{3}(E(x_1) + E(x_3) + E(x_5)) = \mu$ and $E(\overline{\mu}) = \frac{1}{2}(E(x_2) + E(x_4)) = \mu$, so all three are unbiased.*

5. **05 - Week 5**

MULTI | 1.0 point | 0.10 penalty | Single | Shuffle

The multiple regression model in matrix form is

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

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

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

in which $\widehat{\beta}$ is the OLS estimator of $\beta$ and $\widehat{\mathbf{u}}$ is the vector of OLS residuals. We know that columns of $\mathbf{X}$ are linearly independent. What other assumption do we need to be able to say that $\widehat{\beta}$ is an unbiased estimator of $\beta$?

(a) $\mathbf{X}'\widehat{\mathbf{u}} = \mathbf{0}$.
(b) We don't need any other assumption.
(c) $Var(\mathbf{u} \mid \mathbf{X}) = \sigma^2\mathbf{I}_n$.
(d) $E(u_i) = 0$ for $i = 1, 2, \ldots, n$.
(e) $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$. (100%)

*For unbiasedness of the OLS estimator, we need the model to be linear in parameters, columns of the $\mathbf{X}$ matrix to be linearly independent, and $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$. The only one missing is the last one.*

*Total of marks: 5*

**Introductory Econometrics**
Tutorial 5 Solutions

**PART B:**

1. 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%

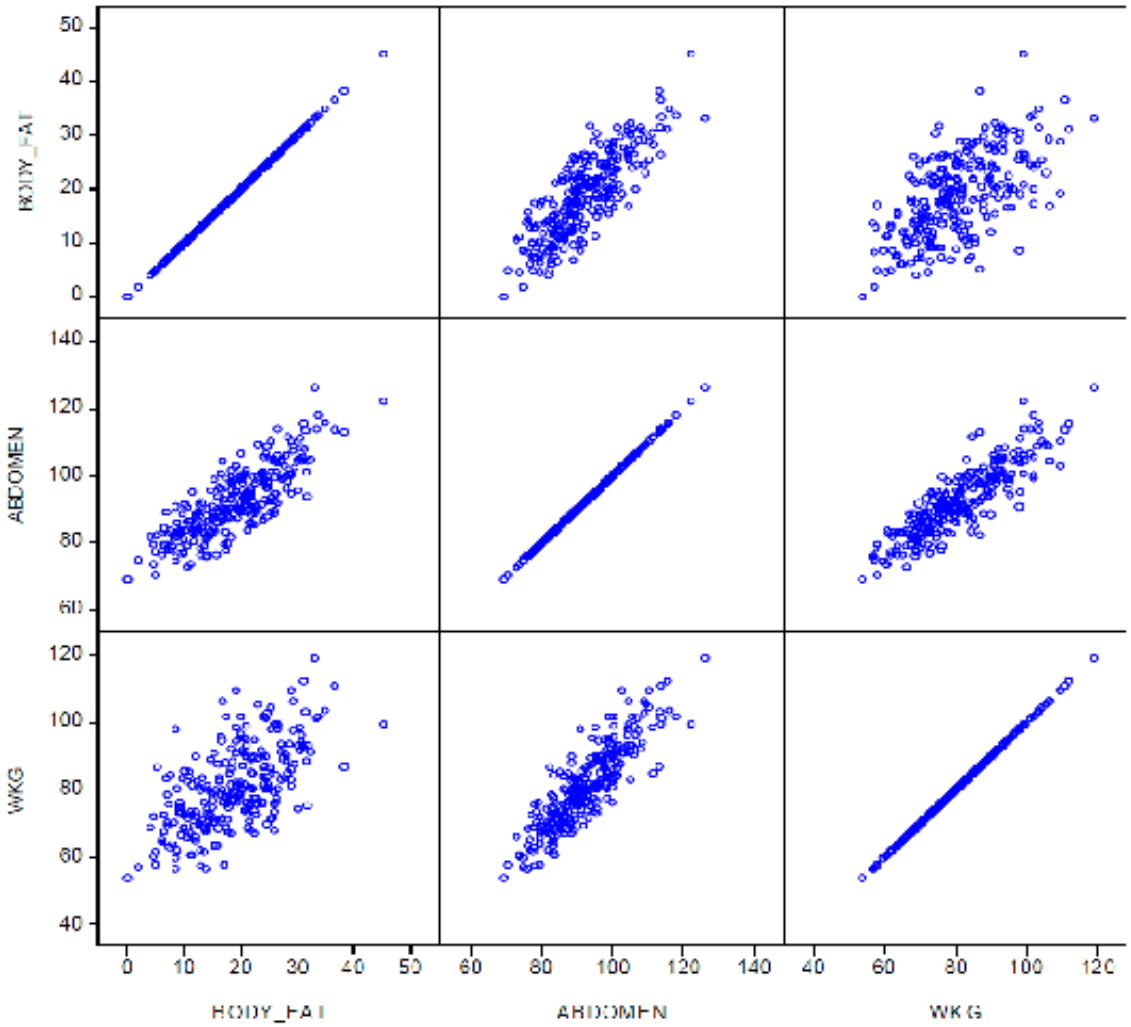
v) The first house in the sample has $sqrft = 2{,}438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

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

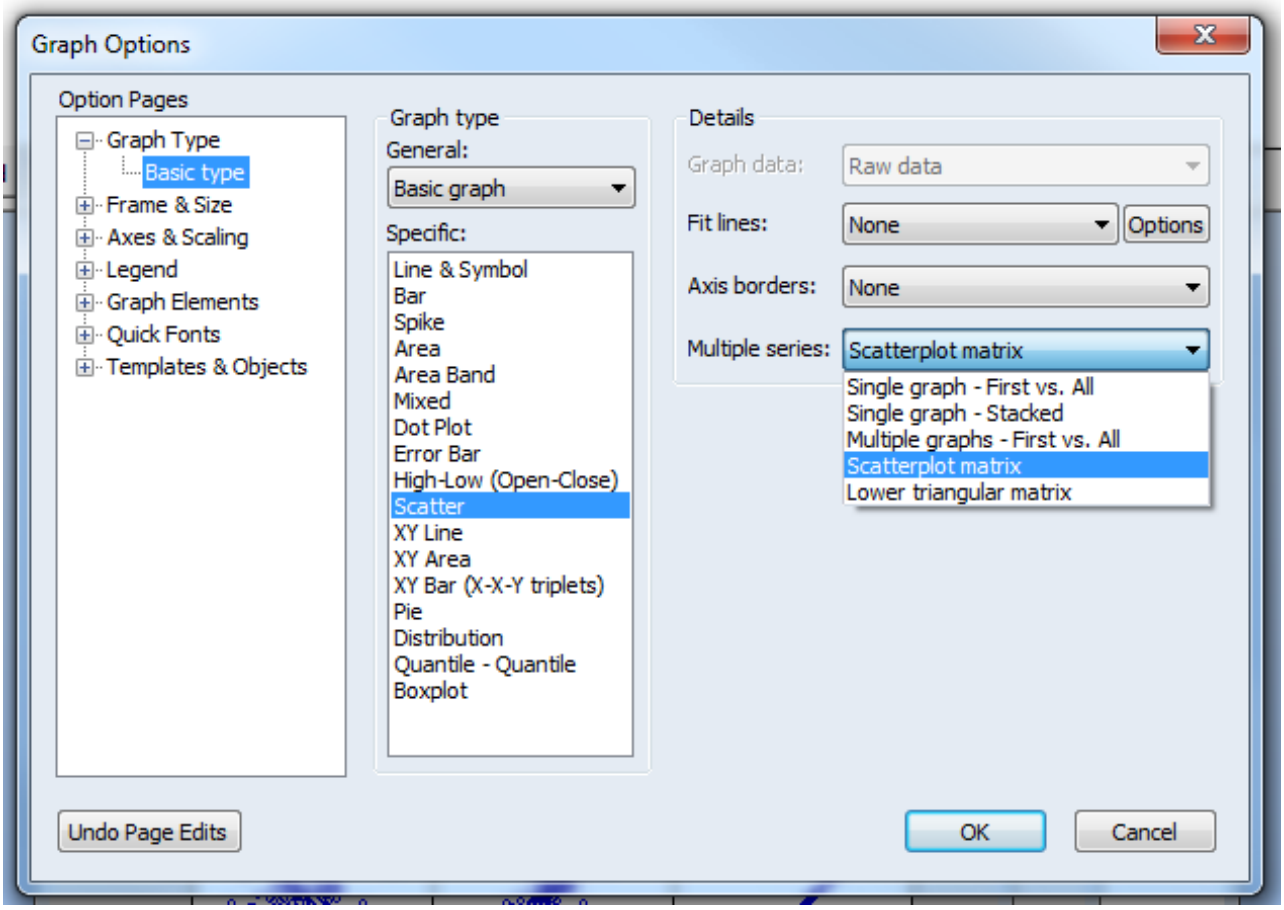
vi) The actual selling price of the first house in the sample was \$300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

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.

2. 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.

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

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

(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 bodyfatkg.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: 1 251
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 |

| | | | |
|---|---|---|---|
| Root MSE | 4.042134 | 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 | 1.786889 | Prob(F-statistic) | 0.000000 |

Dependent Variable: BODY_FAT
Method: Least Squares
Sample: 1 251
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 |

| | | | |
|---|---|---|---|
| Root MSE | 4.042134 | 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 | 1.786889 | Prob(F-statistic) | 0.000000 |