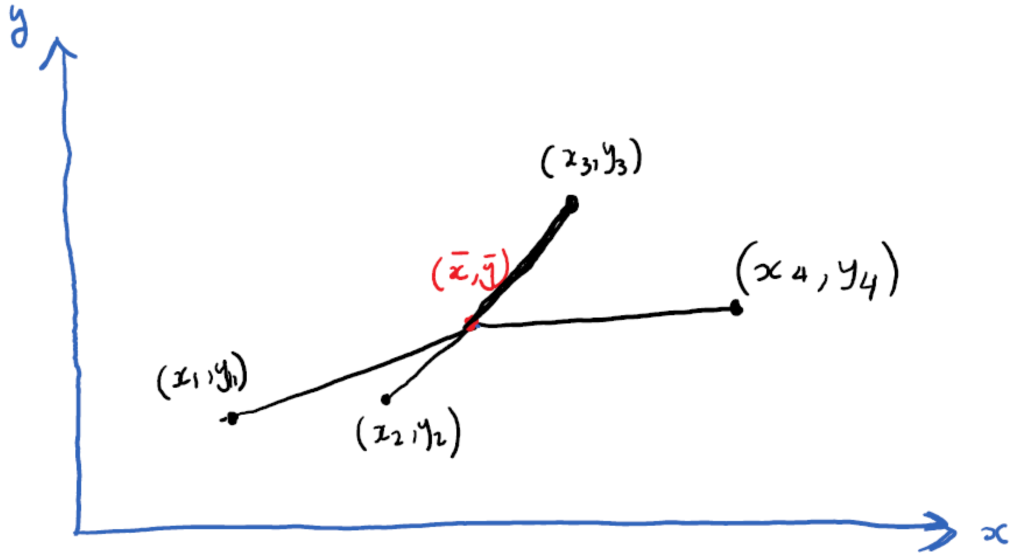
Introductory Econometrics

Tutorial 5 Solutions

PART A: Solutions for Part A will be provided after the assignment due data.

PART B: You do not need to hand this part in. It will be covered in the tutorial. It is still a good idea to attempt these questions before the tutorial.

1.



As you can see, it is not such a crazy idea to say that the average of these four slopes is a reasonable estimator for the slope of the population regression line.

$$\begin{aligned}
\text{slope of line connecting } (x_1, y_1) \text{ to } (\bar{x}, \bar{y}) &= \frac{y_1 - \bar{y}}{x_1 - \bar{x}} \\
\text{slope of line connecting } (x_2, y_2) \text{ to } (\bar{x}, \bar{y}) &= \frac{y_2 - \bar{y}}{x_2 - \bar{x}} \\
\text{slope of line connecting } (x_3, y_3) \text{ to } (\bar{x}, \bar{y}) &= \frac{y_3 - \bar{y}}{x_3 - \bar{x}} \\
\text{slope of line connecting } (x_4, y_4) \text{ to } (\bar{x}, \bar{y}) &= \frac{y_4 - \bar{y}}{x_4 - \bar{x}}
\end{aligned}$$

Therefore $\hat{\beta}_1^{[5]} = \frac{1}{4}\left(\frac{y_1-\bar{y}}{x_1-\bar{x}} + \frac{y_2-\bar{y}}{x_2-\bar{x}} + \frac{y_3-\bar{y}}{x_3-\bar{x}} + \frac{y_4-\bar{y}}{x_4-\bar{x}}\right)$.
Instead of writing a similar expression with different indices four times, we could write

$$\begin{aligned}
\text{slope of line connecting } (x_t, y_t) \text{ to } (\bar{x}, \bar{y}) &= \frac{y_t - \bar{y}}{x_t - \bar{x}} \text{ for } t = 1, \ldots, 4 \\
\text{and use that to write } \hat{\beta}_1^{[5]} &= \frac{1}{4}\sum_{t=1}^{4} \frac{y_t - \bar{y}}{x_t - \bar{x}}
\end{aligned}$$

Generalising this to any number of observations is then achieved simply with replacing 4 by n, i.e. $\hat{\beta}_1^{[5]} = \frac{1}{n}\sum_{t=1}^{n} \frac{y_t-\bar{y}}{x_t-\bar{x}}$.

- The steps for showing unbiasedness of an estimator:

  (a) Use the population relationship to replace y in the estimator formula and see if you can get the true parameter by itself: Here, the estimator involves $\bar{y}$ also. From the population formula $y_t = \beta_0 + \beta_1 x_t + u_t \quad t = 1, \ldots, n$, we can obtain that (if you sense that the summation sign is causing confusion, show the rest of the problem by writing the sum explicitly for $n = 4$ and not use the summation sign).

  $$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$$

  Note that $E(u_t) = 0$ does not imply that $\bar{u} = 0$. (Explain this and make sure everyone understands the difference between sample mean and population mean). However, it does imply that $E(\bar{u}) = 0$. Using the population model to substitute for y_t and $\bar{y}$ in $\hat{\beta}_1^{[5]}$ formula, we get

  $$\begin{aligned} \hat{\beta}_1^{[5]} &= \frac{1}{n}\sum_{t=1}^{n} \frac{y_t - \bar{y}}{x_t - \bar{x}} = \frac{1}{n}\sum_{t=1}^{n} \frac{\beta_0 + \beta_1 x_t + u_t - (\beta_0 + \beta_1 \bar{x} + \bar{u})}{x_t - \bar{x}} \\ &= \frac{1}{n}\sum_{t=1}^{n} \frac{\beta_1 (x_t - \bar{x}) + u_t - \bar{u}}{x_t - \bar{x}} = \frac{1}{n}\sum_{t=1}^{n} \beta_1 + \frac{1}{n}\sum_{t=1}^{n} \frac{u_t - \bar{u}}{x_t - \bar{x}} \\ &= \beta_1 + \frac{1}{n}\sum_{t=1}^{n} \frac{u_t - \bar{u}}{x_t - \bar{x}}. \end{aligned}$$

  (b) Take expectations conditional on the explanatory variables to see if the expected value of the estimator is equal to the parameter of interest:

  $$\begin{aligned} E\left(\hat{\beta}_1^{[5]} \mid \mathbf{X}\right) &= \beta_1 + \frac{1}{n}\sum_{t=1}^{n} E\left(\frac{u_t - \bar{u}}{x_t - \bar{x}} \mid \mathbf{X}\right) \\ &\quad \text{because expectation goes through the summation} \\ &= \beta_1 + \frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{x_t - \bar{x}} E(u_t - \bar{u} \mid \mathbf{X})\right) \\ &\quad \text{because } x_t \text{ and } \bar{x} \text{ can be treated as constants given } \mathbf{X} \\ &= \beta_1 + \frac{1}{n}\sum_{t=1}^{n}\left(\frac{1}{x_t - \bar{x}} \left(E(u_t \mid \mathbf{X}) - E(\bar{u} \mid \mathbf{X})\right)\right) \\ &= \beta_1 \end{aligned}$$

  because $E(\mathbf{u} \mid \mathbf{X}) = \mathbf{0}$ implies that $E(u_t \mid \mathbf{X}) = 0$ for all t which implies also $E(\bar{u} \mid \mathbf{X}) = 0$.

  (c) Since $E\left(\hat{\beta}_1^{[5]} \mid \mathbf{X}\right) = \beta_1$ which is a constant, therefore $E\left(\hat{\beta}_1^{[5]}\right) = \beta_1$, i.e. $\hat{\beta}_1^{[5]}$ is an unbiased estimator of β_1.

  (d) 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?

$$63.2\%$$

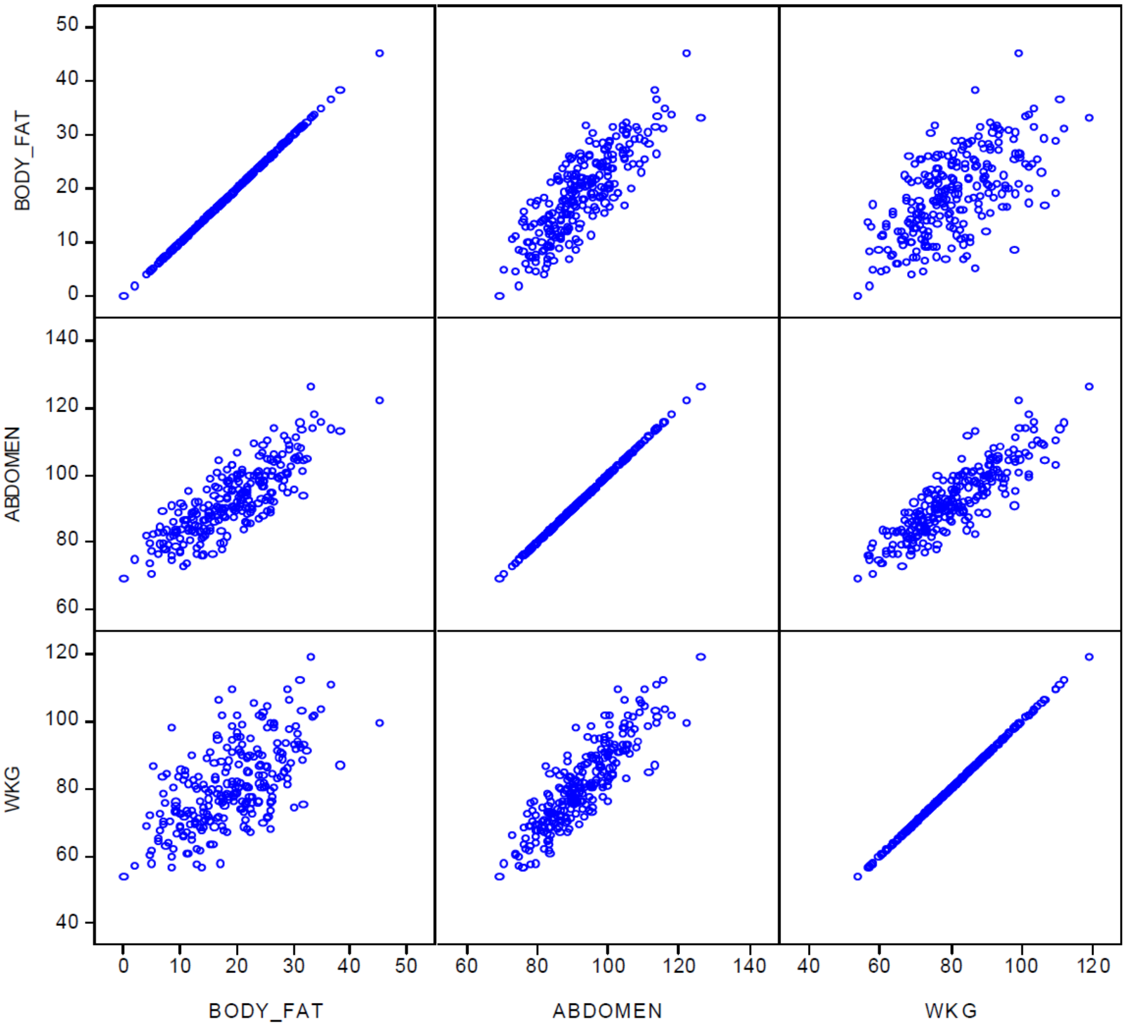
v) The first house in the sample has $sqrft = 2{,}438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

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

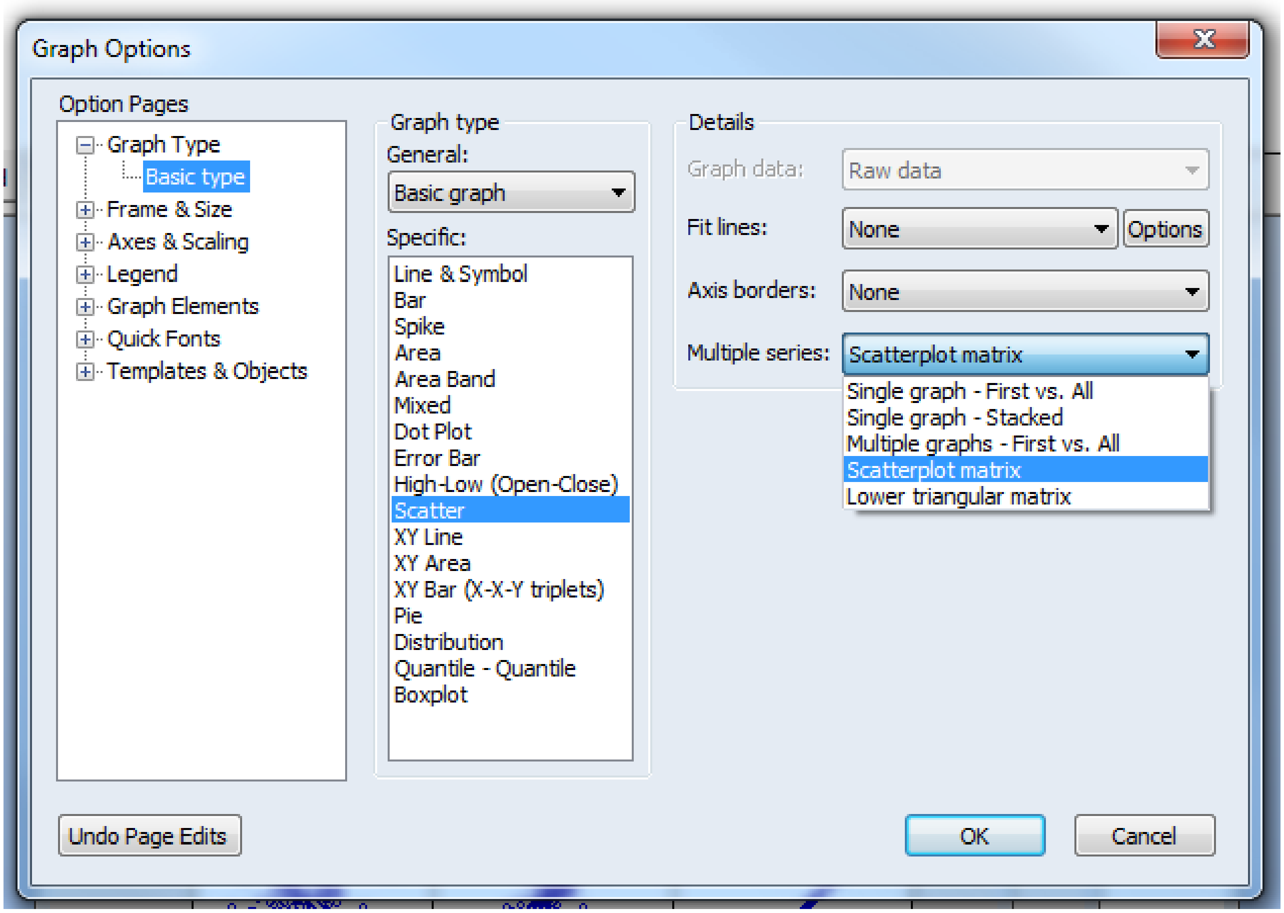
vi) The actual selling price of the first house in the sample was \$300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

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.


Graph Options

Graph type

General: Basic graph

Specific: Scatter

Details

Graph data: Raw data

Fit lines: None

Axis borders: None

Multiple series: Scatterplot matrix

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,

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

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,

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

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^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.

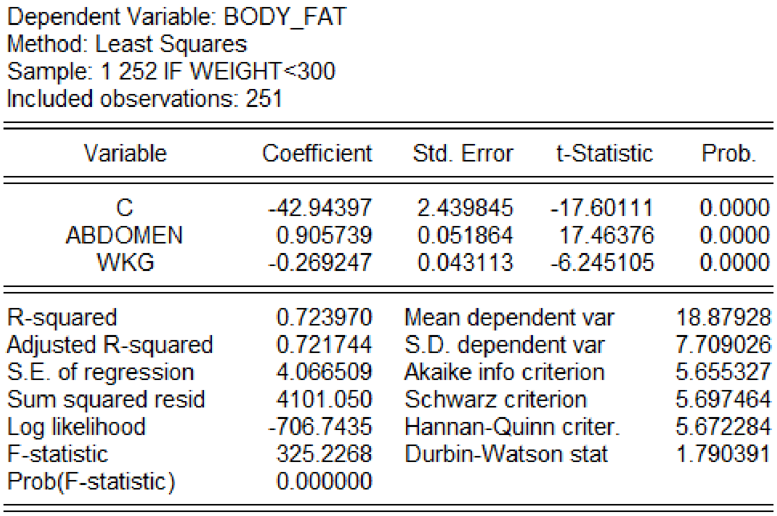
They do :-)

(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

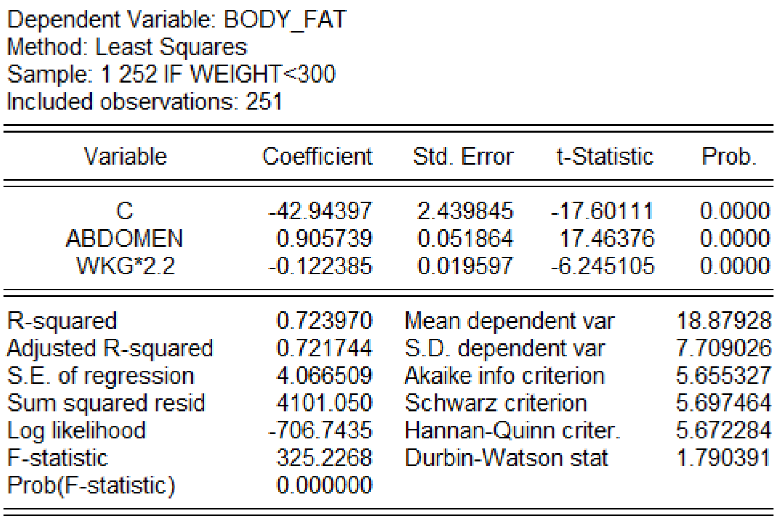


```
Dependent Variable: BODY_FAT
Method: Least Squares
Sample: 1 252 IF WEIGHT<300
Included observations: 251

        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
```

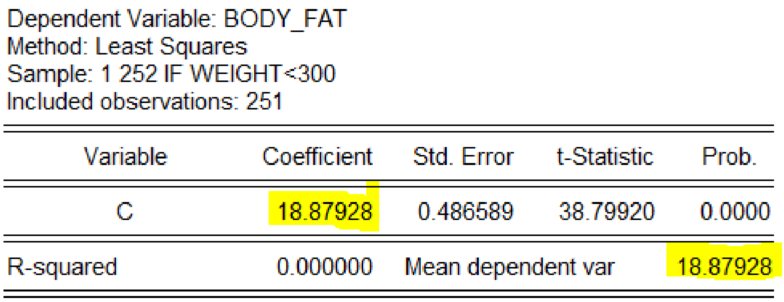


```
Dependent Variable: BODY_FAT
Method: Least Squares
Sample: 1 252 IF WEIGHT<300
Included observations: 251
==========================================================================
        Variable        Coefficient   Std. Error   t-Statistic     Prob.
==========================================================================
           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
==========================================================================
```

(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.

It will be equal to the sample average of body fat.

This was derived in tutorial 4. Ask one to derive it if you feel not everyone has understood.



```
Dependent Variable: BODY_FAT
Method: Least Squares
Sample: 1 252 IF WEIGHT<300
Included observations: 251
==========================================================================
        Variable        Coefficient   Std. Error   t-Statistic     Prob.
==========================================================================
           C              18.87928     0.486589      38.79920     0.0000
==========================================================================
R-squared                 0.000000   Mean dependent var          18.87928
==========================================================================
```