Worksheet for week # 4

1. Suppose that you are a director of university, you are considering to cancel the entrance exams and you search for new criteria how to select good students for your school. Assume you consider to use as the prior criteria the grades from high school (in US called ACT scores).

   In order to decide, some data about current students seem to be helpful. Let's say you have a sample of 15 students. The following table contains the ACT scores and the GPA (grade point average) for these 15 students. Grade point average represents the performance of the students at the university, it is based on a four-point scale and has been rounded to one digit after the decimal.

Student	GPA	ACT
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30
9	3.2	23
10	3.0	28
11	3.5	30
12	2.5	20
13	3.8	32
14	2.6	26
15	2.7	23

   (a) Estimate the relationship between GPA and ACT using OLS; that is, obtain the intercept and slope estimates in the equation GPA $= \beta_0 + \beta_1$ ACT. Use Excel for the computation.

       i. Compute intercept and slope coefficients using the summation formulas for β_0 and β_1.

       ii. Compute intercept and slope coefficients using matrix formula for $\boldsymbol{\beta}$.[1]

   (b) Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the GPA predicted to be, if the ACT score is increased by 5 points?

---

[1]You may want to check out the Excel commands =MMULT() and =MINVERSE().

(c) Find and list the fitted values and the residuals of the model.

(d) Verify that the residuals (approximately) sum to zero.

(e) What is the predicted value of GPA when ACT=20?

2. This exercise serves to illustrate the distinction between the stochastic error term and the residual. Usually, we can never observe the error term, but we can get around this difficulty if we assume values for the true coefficients. Calculate values of the error term and residual for each of the following six observations given that the true β_0 equals 0, the true β_1 equals 1.5.

$$
\begin{array}{ccccccc}
Y_i & 2 & 6 & 3 & 8 & 5 & 4 \\
X_i & 1 & 4 & 2 & 5 & 3 & 4
\end{array}
$$

3. Open the dataset *401ksubs* in Gretl located in (File → Open data → Sample File).

(a) Display the dataset for visual inspection. For this select all variable by pressing $Ctrl + A$, then go to Data → Display values. Explore the following commands in Data tab: Edit values, Add observations and Dataset structure.

(b) Plot the histogram of annual income (right click on inc, then choose Frequency distribution).

(c) Compute descriptive statistics (right click on inc, then Summary statistics).

(d) Compute Gini coefficient for the income data (Variable → Gini coefficient). Interpret your findings.

(e) Plot the histogram of ln(income). Explain how and why it is different from the histogram of income.

4. Now split the sample into several sub-samples.

(a) Compute descriptive statistics and plot histograms for males and females (Sample → Restrict, based on criterion...). Interpret your findings.

(b) Re-do the above exercise for married and unmarried individuals. Explain your intuition.

5. Let's now move to bivariate graphs. Make three scatter plots: income against family size, income against age and financial assets against income. Further, compute correlation coefficients for each pair of variables. Explain your intuition.

6. Now let's add time dimension to our analysis. Open dataset *greene5_1* and make time series plots of variables that you might find interesting.

7. Open dataset *greene14_1* and explore the panel structure of the dataset. Make panel plots of variable Q by company (unit).