Exercise Class 1 - OLS

Instructor: Irene Iodice
Email: irene.iodice@malix.univ-paris1.fr

Ex.1: Derivation of OLS
Derive the formula for the OLS intercept and slope coefficient by minimizing the sum of the squares of the vertical deviations from each data point to the regression line.

Ex.2: Interpreting the HPs behind OLS
Let hwage denote the hourly wage of Italian workers, and let educ be the number of years of education. A simple model relating earnings to education can be:

$$hwage_i = \beta_0 + \beta_1 educ_i + u_i \tag{1}$$

1. *What kinds of factors are contained in u_i? Are these likely to be correlated with level of education?*
2. *Will a simple regression analysis uncover the ceteris paribus effect of education on wage? Explain.*
3. *Suppose that $E(u_i) \neq 0$. Rewrite the model so that the new error has a zero expected value. What has changed?*

Ex.3: Geometrical interpretation of regression line
Our firm 'Pippo' this week hires a consultant to predict the value of weekly sales of their product if their weekly advertising is increased to 600 €per week. The consultant takes a record of how much the firm spent on advertising per week and the corresponding weekly sales over the past 6 months. The consultant writes 'Over the past 6 months the average weekly expenditure on advertising has been 450 €and average weekly sales have been 7500 €. Based on the results of a simple linear regression, I predict sales will be 8500 €if 600 €per week is spent on advertising.'

1. *What is the estimated simple regression used by the consultant to make this prediction?.*
2. *Sketch a graph of the estimated regression line. Locate the average weekly values on the graph.*
3. *Show by means of the geometrical interpretation showed in class that the point of averages $(\overline{x}, \overline{y})$ lies on the estimated regression line.*

Ex.4: From row data to OLS estimates and analysis
The instructor of some college courses shows to his students the following information on the GPA (Italian scale 18-30) and the TOEFL score obtained on analytical writing to get access to university by 8 students at their college level.

Student	1	2	3	4	5	6	7	8
TOEFL	3	2.7	3.5	2.8	3.4	3.0	3.7	3.6
GPA	26	25	27	21	24	25	30	29

1. *Estimate the relationship between TOEFL score and GPA using OLS. Comment the estimates.*
2. *How much higher is the TOEFL predicted to be, if the GPA score is increased by 5 points?* If GPA is 5 points higher, TOEFL predicted increases by $0.1022(5) = 0.511$.
3. *Compute the fitted values and residuals for each observation and verify that the residuals (approximately) sum to zero.*
4. *What is the predicted value of TOEFL when $GPA = 20$?*
5. *Assuming that the error terms are homoskedastic. Test whether the effect of college results (GPA) on the performance on the TOEFL test is statistically significant.*

6. *What changes if you are not willing to assume homoskedasticity in the errors?*

7. *How much of the variation in TOEFL for these 8 students is explained by GPA? Explain.*

Ex.5: Application
You have the results of a simple linear regression based on provinces level data in Italy with a total of $n = 52$ observations.

1. *The estimated error variance is* $s^2_{\hat{u}} = 2$. *What is the residuals sum of squares?*

2. *The estimated variance of* $\hat{\beta}_1$ *is 0.0016. What is the standard error of* $\hat{\beta}_1$*? What is the value of* $\sum(x_i - \overline{x})^2$ *(assume hp5)?*

3. *Suppose the dependent variable* y_i *is the province's mean income (in thousands of €) of woman older than 18 and* x_i *the share of woman* > 18 *years old that holds an high school diploma. If* $\hat{\beta}_1 = 0.12$, *interpret this result.*

4. *Suppose* $\overline{x} = 80$ *and* $\overline{y} = 19.6$, *what is the estimate of the intercept parameter?*

5. *Given the results in (2) and (4), what is* $\sum_i x_i^2$*?*

6. *What is the standard error of* $\hat{\beta}_0$*?*

7. *For the very nice province of Vicenza the value of* $y_i = 22$ *and the value of* $x_i = 90$. *Compute the least squares residual for Vicenza.*

Ex.6: Heteroskedasticity
Consider the consumption function:

$$cons_i = \beta_0 + \beta_1 inc_i + u_i \qquad u_i = e_i\sqrt{inc_i} \tag{2}$$

where inc_i *is household income,* e_i *is a random variable with* $E(e_i) = 0$ *and* $Var(e) = \sigma^2_e$. *Assume that* e_i *is independent of* inc_i.

1. *Show that HP3 is satisfied, which is that the zero conditional mean assumption holds.*

2. *Show that HP5 is violated, which is that is the variance of the error is the same for any value of the regressor.*

3. *Discuss why one should expect the variance of consumption to increases with family income.*