Exercise Class - Econometrics Class 1

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.

The problem is:

$$min_{\hat{\beta}_0,\hat{\beta}_1} \sum_i^n (y_i - \hat{y}_i)^2 \tag{1}$$

with $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. We solve this minimization problem by equating the partial derivatives of the above function, that we call L, to 0:

$$\begin{cases} \frac{\partial L}{\partial \hat{\beta}_0} = \sum_i^n -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial L}{\partial \hat{\beta}_1} = \sum_i^n -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

Now we will solve for the parameters of interest using some algebra tricks and some properties of summations. Let's start with the first foc: we leave out the -2 and we make use of the fact that $\sum_i^n y_i = n\overline{y}$ to rewrite it as $\hat{\beta}_0 n = \overline{y}n - \hat{\beta}_1 n\overline{x}$. We then get rid of n and obtain the classic formula,

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}. \tag{2}$$

We then handle the second foc, solving for $\hat{\beta}_1$. As before leave out -2, and substitute $\hat{\beta}_0$ in the equation:

$$\sum_i^n (x_i y_i - (\overline{y} - \hat{\beta}_1 \overline{x})x_i - \hat{\beta}_1 x_i^2) = 0 \tag{3}$$

Then split the summation and take the averages (constant terms) out of them:

$$\sum_i^n x_i y_i - \overline{y}\sum_i^n x_i + \hat{\beta}_1 \overline{x}\sum_i^n x_i - \hat{\beta}_1 \sum_i^n x_i^2 = 0 \tag{4}$$

$$\hat{\beta}_1 = \frac{\sum_i^n x_i y_i - \overline{y}\sum_i^n x_i}{\sum_i^n x_i^2 - \overline{x}\sum_i^n x_i} = \frac{\sum_i^n x_i y_i - n\overline{xy}}{\sum_i^n x_i^2 - n\overline{x}^2} \tag{5}$$

Note that this equation is equivalent to the one Pr. Secchi showed you in class:

$$\hat{\beta}_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} \tag{6}$$

since $\sum_i^n (x_i - \overline{x})(y_i - \overline{y}) = \sum_i^n x_i y_i - n\overline{xy}$ and $\sum_i^n (x_i - \overline{x})^2 = \sum_i^n x_i^2 - n\overline{x}^2$.

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:

$$hwage_i = \beta_0 + \beta_1 educ_i + u_i \tag{7}$$

1. *What kinds of factors are contained in u_i? Are these likely to be correlated with level of education?*
   It might contain an error due to the wrong functional form (es. the quadratic term). Variables that are left out for principle of parsimony: they explain hourly wage but not education. If they are correlated to education they become omitted variables. Geographical location, the number of years of experience, sex, family background as family income and age are just a few possibilities. For example, family income and education are probably positively correlated; age and education may be negatively correlated because in more recent cohorts have, on average, more years of educ, etc.

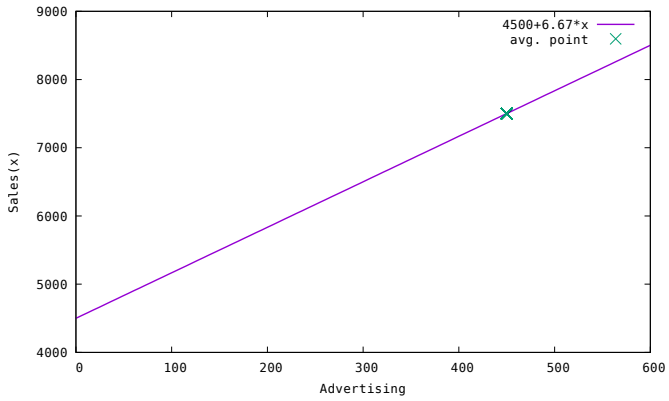
2. *Will the model above provide a good estimate of the effect of education on wage? Explain.*
   Not if the factors we listed above are correlated with *educ*. If u_i is correlated with $educ_i$ then $E(u_i|educ_i) \neq 0$, and so HP3 fails and then β_1 will be biased and inconsistent. This means that it does not summarizing the real impact of education on wage, as it includes also all the co-founding factors that correlates with *educ*.

3. *Suppose that $E(u_i) \neq 0$. Rewrite the model so that the new error has a zero expected value. What has changed?*
   Let denote $E(u_i) = \alpha$; in the equation $hwage_i = \beta_0 + \beta_1 educ_i + u_i$, add and subtract α from the right hand side to get $hwage_i = (\alpha + \beta_0) + \beta_1 educ_i + (u_i - \alpha)$. Call the new error $e_i = u_i - \alpha$, so that $E(e_i) = 0$. The new intercept is $(\alpha + \beta_0)$, but the slope is still β_1.

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?*
   $\begin{cases} \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 7500 - \hat{\beta}_1 450 \\ 1000 = \hat{\beta}_0 + \hat{\beta}_1 600 \end{cases}$ So that $\hat{\beta}_1 = \frac{1000}{150}$ and $\hat{\beta}_0 = 4500$.

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 (x,y) lies on the estimated regression line..*
   Recall that

   - $\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{(x_i - \overline{x})^2} = \frac{\sum_i (x_i - \overline{x})^2 \frac{y_i - \overline{y}}{x_i - \overline{x}}}{\sum_i (x_i - \overline{x})^2} = \sum_i w_i \frac{(y_i - \overline{y})}{(x_i - \overline{x})}$ with $w_i = \frac{(x_i - \overline{x})^2}{\sum_i (x_i - \overline{x})^2}$
     This means that $\hat{\beta}_1$ is a weighted average of the slopes of the lines linking each point (x_i, y_i) with the mean point $(\overline{x}, \overline{y})$.

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
TOEFLaw	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.*
We want to estimate $\hat{TOEFL} = \hat{\beta}_0 + \hat{\beta}_1 GPA$. Recall from before that:

   - $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$
   - $\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{\sum_i x_i y_i - n\overline{xy}}{\sum_i x_i^2 - n\overline{x}^2}$

   In our case we have $y_i = TOEFL_i$, $x_i = GPA_i$, and $n = 8$. The table below will provide all the info that we use in solving the questions: $\overline{y} = 25.7/8 = 3.2125$ and $\overline{x} = 207/8 = 25.875$, and obtain the numerator as $\sum_i x_i y_i - n\overline{xy} = 670.8 - 8 \times 3.2125 \times 25.875 = 5.8125$ and denominator $\sum_i (x_i)^2 - n\overline{x}^2 = 5413 - 5356.125 = 56.875$. We obtain the slope as $\hat{\beta}_1 = 5.8125/56.875 \approx 0.1022$. Then, we can write:

$$TOEFL = 0.5681 + 0.1022 GPA, n = 8. \tag{8}$$

   The intercept does not have a useful interpretation because the the GPA is not close to zero for the population of interest.

i	y_i	x_i	$y_i x_i$	x_i^2	y_i^2	$\hat{y}_i$	$\hat{u}_i$	$\hat{u}_i^2$	$(x_i - \overline{x})^2 \hat{u}^2$
1	3	26	78	676	9	3.2253	-0.2253	0.048	0.0008
2	2.7	25	67.5	625	7.29	3.1231	-0.4231	0.179	0.137
3	3.5	27	94.5	729	12.25	3.3275	0.1725	0.03	0.038
4	2.8	21	58.8	441	7.84	2.7143	0.0857	0.007	0.1664
5	3.4	24	81.6	576	11.56	3.0209	0.3791	0.1437	0.5052
6	3.0	25	75	625	9	3.1231	-0.1231	0.015	0.0115
7	3.7	30	111	900	13.69	3.6342	0.0659	0.004	0.068
8	3.6	29	104.4	841	12.96	3.5319	0.0681	0.004	0.039
$\sum_i$	25.7	207	670.8	5431	83.59		-0.002	0.4323	0.9659

2. *How much higher is the TOEFL score predicted to be, if the GPA score is increased by 5 points?*
If GPA is 5 points higher, TOEFL predicted increases by $O.1022(5) = 0.511$.

3. *Compute the fitted values and residuals for each observation and verify that the residuals (approximately) sum to zero.*

   - $i = 1$, $\hat{y} = 0.5681 - 0.1022 \times 26 = 3.2253$, then the residual is $\hat{u}_i = y_i - \hat{y}_i = 3 - 3.2253 = -0.2253$
   - for $i > 1$ compare the res. with the table

   As reported in the table, the residuals sum up to -0.0002, which is pretty close to zero given the inherent rounding error.

4. *What is the predicted value of TOEFL when $GPA = 20$?*
When $GPA = 20$ we have $\hat{TOEFL} = 0.5681 + 0.1022(20) \approx 2.61$.

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.*
Remember that to for the t-statistic we need to standardize the estimator by its mean and standard deviation under the null hypothesis. Recall also that

$$SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \frac{\frac{1}{n-2} \sum_i (x_i - \overline{x})^2 \hat{u}_i^2}{\frac{1}{n} \left[\sum_i (x_i - \overline{x})^2 \right]^2}} \tag{9}$$

under homoskedasticity robust standard errors are (asymptotically) identical and thus $\sum_i (x_i - \overline{x})^2 u_i^2 = \sum_i (x_i - \overline{x})^2 \sum_i \hat{u}^2$. Then the above formula reduces to:

$$SE(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{\frac{1}{n}\frac{s^2_{\hat{u}}}{\sigma^2_x}} \text{ where } s^2_{\hat{u}} = \frac{1}{n-2}\sum_i \hat{u}_i^2 = \frac{SSR}{n-2} \text{ and } \sigma^2_x = \frac{1}{n}\sum_i (x_i - \overline{x})^2$$

In our case we have:

- $s^2_{\hat{u}} = \frac{0.4323}{6} = 0.07$
- $s^2_x = \frac{\sum_i (x_i)^2 - n\overline{x}^2}{n} = \frac{5413 - 5356.125}{8} = \frac{56.875}{8} = 7.1.$
- $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{8}\frac{0.07}{7.1} = 0.001$

We run the following test:

- Assumptions:
  - The HP we made to run OLS hold true (HP1-4)
  - We assume HP5 as suggested;
  - We must assume HP6 on normal distribution the error since we are running inference with a small sample.
- Hypotheses:
  $H_0: \quad \beta_1 = 0$
  $H_1 \quad \beta_1 \neq 0$
- Test statistic, given H_0 is true, is:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} =$$

- Distribution of test statistic:
  If the assumptions are met and H_0 is true, the test statistic is distributed as Student's t distribution with 6 (which is n-2, 2 df less from estimating both coefficients β_0 and β_1), degrees of freedom.
- Testing procedure:
  - Decision rule: with $\alpha = .05$ and $df = 6$, the test is two tails, the critical value of $t_{\alpha/2,7} = 2.45$. We reject H_0 if $|t^{act}| > 2.45$.
  - Calculation of test statistic:

$$t^{act} = \frac{0.1022 - 0}{\sqrt{0.001}} = \frac{0.1022}{0.032} = 3.05 \tag{10}$$

  - Statistical decision: Since $t^{act} = 3.05 > t = 2.45$, we reject H_0 and conclude at 0.05 level of significance that GPA has a significant effect on TOEFL.

6. *What changes if you are not willing to assume homoskedasticity in the errors?* We compute the heteroskedasticity-robust standard errors through the following formula and by means of the values computed in the table:

- $SE(\hat{\beta}_1) = \sqrt{\frac{\frac{1}{n}\frac{1}{n-2}\sum_i (x_i - \overline{x})^2 \hat{u}_i^2}{\frac{1}{n}\left[\sum_i (x_i - \overline{x})^2\right]^2}} = \sqrt{\frac{\frac{1}{8}\frac{1}{6}0.9659}{\frac{1}{8}(0.4323)^2}} = \sqrt{0.86}$

IN this case the test statistic:

$$t^{act} = \frac{0.1022 - 0}{\sqrt{0.86}} = \frac{0.1022}{0.93} = 0.11 \tag{11}$$

So in this case we fail to reject H_0. Note that indeed robust standard errors are typically larger than non-robust standard errors, so the practice can be viewed as an effort to be conservative (from Angrist and Pischke).

7. *How much of the variation in TOEFL for these 8 students is explained by GPA? Explain.*
The sum of squared residuals, $\sum_i \hat{u}_i^2 = 0.4323$ and the total sum of squares, $\sum_i (y_i - \overline{y})^2 = \sum_i (y_i)^2 - n\overline{y}^2 = 1.0288$. So the R-squared from the regression is $R^2 = 1 - SSR/SST \approx 1 - (.4323/1.0288) \approx 0.579$. Therefore, about 58% of the variation in TOEFL is explained by GPA in this small sample of students.

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 $s_{\hat{u}}^2 = 2$. What is the sum of the squared least squares residuals?*
Recall that $s_{\hat{u}}^2 = \frac{\sum \hat{u}^2}{n-2}$, then the RSS (residuals sum of sq.) is $\sum \hat{u}^2 = s_{\hat{u}}^2(n-2) = 2 \times 50 = 100$

2. *The estimated variance of $\hat{\beta}_2$ is 0.0016. What is the standard error of $\hat{\beta}_2$? What is the value of $\sum (x_i - \overline{x})^2$ (assume hp5)?*
We have
   - $SE(\hat{\beta}_2) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}} = \sqrt{0.0016} = 0.04$
   - $\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n}\frac{s_{\hat{u}}^2}{\sigma_x^2}$ and $\sigma_x^2 = \frac{1}{n}\sum_i^n (x_i - \overline{x})^2$ then $\sum_i^n (x_i - \overline{x})^2 = \frac{s_{\hat{u}}^2}{\hat{\sigma}^2_{\hat{\beta}_1}} = \frac{2}{0.0016} = 1250$

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 $\hat{\beta}_2 = 0.12$, interpret this result.*
The estimate for β_2 suggest that a 1% increase in the percentage of woman > 18 who hold high school graduates is associated to an average increase of 120 €in their mean income.

4. *Suppose $\overline{x} = 80$ and $\overline{y} = 19.6$, what is the estimate of the intercept parameter?*
Recall $\beta_1 = \overline{y} - \beta_2 \overline{x} = 19.6 - 0.12 * 80 = 10$

5. *Given the results in (2) and (4), what is $\sum_i x_i^2$?*
Since $\sum_i^n (x_i - \overline{x})^2 = \sum_i^n x_i^2 - n\overline{x}^2$ we have
   - $\sum_i^n x_i^2 = \sum_i^n (x_i - \overline{x})^2 + n\overline{x}^2 = 1250 + 52(80)^2 = 334,050$

6. *What is the standard error of $\hat{\beta}_0$?*
   - $SE(\hat{\beta}_0) = \sqrt{\frac{1}{n}\frac{s_{\hat{u}}^2}{\sigma_x^2} E[x_i^2]}$

   we can compute $E[x_i^2]$ from the results we obtained above as $E[x_i^2] = \frac{1}{n}\sum_i^n x^2 = 334,050/52 \approx 6424$, then
   - $SE(\hat{\beta}_0) = \sqrt{0.0016 \times 6424} = 3.20$

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.*
For the very nice town of Vicenza:
   - $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 22 - 10 - 0.12 \times 90 = 1.2$

Ex.6: Heteroskedasticity
Consider the consumption function:

$$cons_i = \beta_0 + \beta_1 inc_i + u_i \qquad u_i = e_i\sqrt{inc_i} \tag{12}$$

where e is a random variable with $E(e_i) = 0$ and $Var(e) = \sigma_e^2$. Assume that e_i is independent of inc_i.

1. *Show that HP3 is satisfied, which is that the zero conditional mean assumption holds.*
   Recall that HP3 states that $E(u_i|x_i) = 0$. In our case we have $E(e_i\sqrt{inc_i}|inc_i) = \sqrt{inc_i}E(e_i|inc_i) = \sqrt{inc_i} \times 0 = 0$, thus it is because (i) when we condition on the variable inc_i in computing an expectation, its value becomes a constant, (ii) e_i is independent of inc_i thus its conditional expectation is equal to the unconditional one, which in turn is equal zero by Ass.

2. *Show that HP5 is violated, which is that the error terms and the regressor are not independent.*
   Again, when we condition on inc_i in computing a variance, whatever function of inc becomes a constant. So $Var(u_i|inc_i) = Var(e_i\sqrt{inc_i}|inc_i) = (\sqrt{inc_i})^2 Var(e_i|inc_i) = inc_i\sigma_e^2$ because $Var(e_i|inc_i) = \sigma_e^2$.

3. *Discuss why one should expect the variance of consumption to increases with family income.*
   Families with low incomes do not have much discretion about consumption, in the sense that their income is constrained to the necessary goods as housing, food, schooling, and other necessities. Higher income people have more discretion, and some might choose more consumption while others more saving. This discretion suggests wider variability in consumption among higher income families.

4. *How do you interpret the intercept of this model?*
   The intercept gives us the average consumption of people with no income ($inc_i = 0$), we therefore expect it to be positive.