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:

$$min_{\beta_0,\beta_1} \sum_i^n (y_i - \beta_0 - \beta_1 x_i)^2 \tag{1}$$

We solve this minimization problem by equating the partial derivatives of the above function, that we call L, to 0:

$$\begin{cases} \frac{\partial L}{\partial \beta_0} = 0 & \sum_i^n -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \frac{\partial L}{\partial \beta_1} = 0 & \sum_i^n -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \end{cases}$$

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:

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

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

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 $\sum_i^n (x_i - \overline{x})(y_i - \overline{y}) = \sum_i^n x_i y_i - n\overline{x}\overline{y}$ 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?*
   Geographical location, the number of years of experience, sex, family background as family income and age are just a few possibilities. It seems that each of these could be correlated with years of education. 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 a simple regression analysis uncover the ceteris paribus effect of education on wage? Explain.*

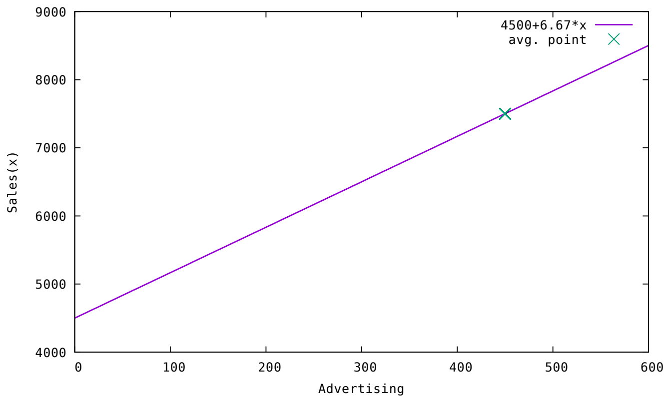
Not if the factors we listed above are correlated with *educ*. Because we would like to hold these factors fixed, they are part of the error term. But if u_i is correlated with $educ_i$ then $E(u_i|educ_i) \neq 0$, and so HP3 fails and then β_1 will be biased and inconsistent. This means that it does 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 $E(u_i) \neq 0$, but is constant. 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?*
   $\begin{cases} \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 7500 - \hat{\beta}_1 450 \\ 8500 = \hat{\beta}_0 + \hat{\beta}_1 600 \end{cases}$ 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 at that the point of averages (x,y) lies on the estimated regression line..*