Introduction to Econometrics
PS #4 - Introduction to Simple Regression Model

1 Population and sample regression function

Suppose you are interested in the relationship between class attendance (X) and test scores (Y) for your cohort.

1. Assume that you have data for the entire class (population). How would you best describe the average relationship denoted by $E(Y_i|X_i)$?
2. The $E(Y_i|X_i)$ (also called the CEF) does not fit all the data points perfectly. Write the equation for Y_i that accounts for this error.
3. Getting data on the entire population is often costly, suppose you draw a random sample instead. What is the best way to describe the relationship from your sample?

2 Derivation of OLS estimators

The true relationship in the population is: $Y_i = \beta_0 + \beta_1 X_i + u_i$. However, with access to only a random sample, we need the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

1. Derive the OLS estimators by minimizing the sum of squared distances between each sample data point and the SRF.
2. Show that $\hat{\beta}_1 = \dfrac{cov(X_i, Y_i)}{V(X_i)}$. Explain the intuition behind this estimator.

3 Re-scaling variables

Your boss asks you to estimate the relationship between the factory's output (Y_i) and the distance of the factory from the headquarters (D_i). Assume that the data on distance (D_i) is originally measured in Kilometers. However, your boss wants you to estimate this relationship using distance measured in Meters.

1. Specify the original regression model. Report the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.
2. Specify the regression model of interest by scaling D_i appropriately. [Note: 1km = 1000m]
3. Find the OLS estimators $\tilde{\beta}_0$ and $\tilde{\beta}_1$.
4. Your boss is highly indecisive - he now asks for the relationship with distance measured in Kilometers. Find the resulting OLS estimate $\hat{\beta}_1$.

Suppose that you multiplied both the output (Y_i) and the distance (D_i) (measured in Kilometers) by 1000

5. Specify the regression model by scaling both variables.
6. Find the OLS estimators, are the results equivalent to $\mathcal{Q}3.1$?

4 Slight Detour: Law of Iterated Expectations

Assume that X and Y are two jointly distributed discrete random variables. Prove that $E(Y) = E\Big\{E(Y|X)\Big\}$

[**Note**: We have already seen LIE in action, in Problem Set #2 $\mathcal{Q}$.3]

5 Gauss-Markov Theorem

Suppose that we are interested in the relationship between wages and the years of education for France. The simple regression model for the same is specified as:

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

1. Is the model linear in parameters?
2. Suppose that our sample is only drawn from the players of Paris Saint-Germain F.C. Would you consider this as a random sample in the context of the entire French labor market?
3. Assuming that $V(educ_i) = 0$, find the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Do the estimators give any information on the relationship of interest?
4. Show that
$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(educ_i - \overline{educ})u_i}{\sum_{i=1}^{n}(educ_i - \overline{educ})^2}$$
Use the ZCM assumption *i.e.*, $E(u_i|e_i) = 0$ to show that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators.
5. The errors of the regression are said to be homoskedastic when $V(u_i|e_i) = 0$. From Figure 1 below, what can you conclude about the variance of u_i as e_i increases?

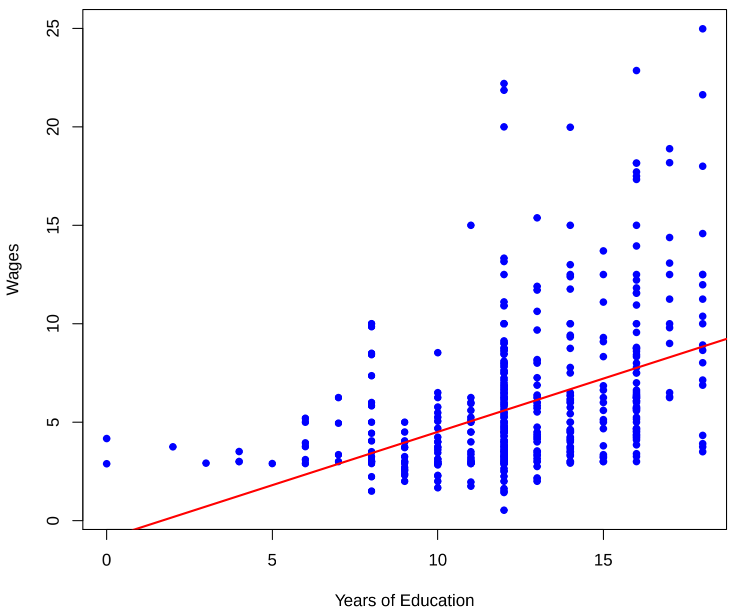

Figure 1: Scatterplot of wages and education + estimated regression line