Ordinary Least Squares estimator

658EC Intermediate Econometrics (Giovanni Millo, DEAMS - University of Trieste, 2025) ©Yves Croissant (2025), Routledge

Simple linear regression model

Obviously there will be rarely a deterministic relationship between a response and one or several covariates. Education has a positive causal effect on wage, but there are many other variables that affect wages: sex, ethnicity, experience, abilities, to name a few. Even if a large set of relevant covariates are observed, some are not (this is in particular the case for abilities). Therefore, we won't try to modelize the value of y as a function of covariates x and some unknown parameters γ: $y = f(x, \gamma)$ but, more modestly, the **conditional expectation** of y:

$$\mathrm{E}(y|x) = f(x, \gamma)$$

Computation of the OLS estimator 1

The model we seek to estimate is: $\mathsf{E}(y_n \mid x_n) = \alpha + \beta x_n$. The difference between the observed value of y and its conditional expectation is called the **error** for observation n:

$$y_n - \mathsf{E}(y_n \mid x_n) = \epsilon_n$$

The linear regression model can therefore be rewritten as:

$$y_n = \mathsf{E}(y_n \mid x_n) + \epsilon_n = \alpha + \beta x_n + \epsilon_n$$

Errors vs. Residuals

ϵ_n is the error for observation n when the values of the unknown parameters (α, β) are set to their true values (α_o, β_o). For given values of (α, β), obtained using an estimation method, ϵ_n will be called the **residual** for observation n. The residual for an observation is therefore the vertical distance between the point for observation n and the regression line.

We seek to draw a straight line that is as closest as possible to all the points of our sample. In the simple linear regression model, the distance between a point and the line is defined by the vertical distance, which is the residual for this observation. For the whole sample, we need to aggregate this individual measures of distance. Summing them is not an issue, as there are positive and negative values of the residuals and the sum may be very close to zero even if the individual residuals are very high in absolute values. One solution would be to use the sum of the absolute values of the residuals, but with the OLS estimator, we'll consider the sum of the squares of the residuals (also called the residual sum of squares (**RSS**).

The "Least Squares" principle

Taking the squares, as taking the absolute values, removes the sign of the individual residuals and it results in an estimator which has nice mathematical and statistical properties. We'll therefore consider a function f which depends on the value of the response and the covariate in the sample (two vectors x and y of length N) and on two unknown parameters α and β, respectively the intercept and the slope of the regression line:

$$f(\alpha, \beta | x, y) = \sum_{n=1}^{N} (y_n - \alpha - \beta x_n)^2$$

Note that we write f as a function of the two unknown parameters conditional on the values of x and y for a given sample.

First-order conditions

First-order conditions for the minimization of f are:

$$\begin{cases} \dfrac{\partial f}{\partial \alpha} &= -2\sum_{n=1}^{N}(y_n - \alpha - \beta x_n) = 0 \\ \dfrac{\partial f}{\partial \beta} &= -2\sum_{n=1}^{N} x_n (y_n - \alpha - \beta x_n) = 0 \end{cases} \tag{1}$$

Or, dividing by -2:

$$\sum_{n=1}^{N}(y_n - \alpha - \beta x_n) = \sum_{n=1}^{N}\epsilon_n = 0 \tag{2}$$

$$\sum_{n=1}^{N} x_n (y_n - \alpha - \beta x_n) = \sum_{n=1}^{N} x_n \epsilon_n = 0 \tag{3}$$

Computing the OLS estimator

Equation 2 indicates that the sum (or the mean) of the residuals in the sample is 0. Dividing this expression by N also implies that, denoting $\bar{y}$ and $\bar{x}$ the sample means of the response and of the covariate:

$$\bar{y} = \alpha + \beta \bar{x},$$

which means that the sample mean is on the regression line.
Denoting $\hat{\epsilon}_n$ the residuals of the OLS estimator, Equation 3 states that $\sum_n x_n \hat{\epsilon}/N = 0$, i.e., that the average cross-product of the covariate and the residual is 0. But, as the sample mean of the residuals $\bar{\hat{\epsilon}}$ is 0, this expression is also the covariance between the covariate and the residuals:

$$\hat{\sigma}_{x\hat{\epsilon}} = \frac{\sum_{n=1}^{N}(x_n - \bar{x})(\hat{\epsilon}_n - \bar{\hat{\epsilon}})}{N} = \frac{\sum_{n=1}^{N} x_n \hat{\epsilon}_n}{N} - \bar{x}\bar{\hat{\epsilon}} = \frac{\sum_{n=1}^{N} x_n \hat{\epsilon}_n}{N} = 0$$

which means that the regression line is such that there is no correlation between the covariate and the residuals in the sample.

Computing the OLS estimator

Subtracting $\bar{y} - \alpha - \beta\bar{x}$, which is 0 from Equation 3, one gets:

$$\sum_{n=1}^{N} x_n \left[(y_n - \bar{y}) - \beta (x_n - \bar{x}) \right] = 0$$

Moreover, $\sum_{n=1}^{N} \bar{x} \left[(y_n - \bar{y}) - \beta (x_n - \bar{x}) \right] = 0$ and so:

$$\sum_{n=1}^{N} (x_n - \bar{x}) \left[(y_n - \bar{y}) - \beta (x_n - \bar{x}) \right] = 0 \tag{4}$$

Solving for β, we finally get the estimator of the slope:

$$\hat{\beta} = \frac{\sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})}{\sum_{n=1}^{N} (x_n - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} = \hat{\rho}_{xy} \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \tag{5}$$

Alternative formulations for the OLS estimator

Equation 5 gives three formulations for this estimator:

- the first indicates that it is the ratio of the covariation of x and y: $S_{xy} = \sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y})$ and the variation of x: $S_{xx} = \sum_{n=1}^{N}(x_n - \bar{x})^2$,
- the second is obtained by dividing both sides of the ratio by the sample size, so that the estimator is now the ratio of the sample covariance between x and y and the sample variance of x,
- the third is obtained by introducing the coefficient of correlation between x and y: $\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$, so that the estimator is also expressed as the product of the coefficient of correlation and the ratio of the standard deviations.

This last formulation is particularly intuitive: $\hat{\rho}_{xy}$ is a pure measure of the correlation between the covariate and the response. The ratio of the standard deviations gives the relevant unit to the slope, which is the unit of y divided by the unit of x.