# [STAT-17] The linear regression model

**Miguel-Angel Canela**
**Associate Professor, IESE Business School**

## The linear regression model

The **linear regression model** is usually presented as a **regression equation**,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon.$$

This probably looks familiar to you. Nevertheless, the linear regression model is not that simple, so it is worth to spend a little time commenting the different elements and assumptions involved. The elements of the model are:

- The **dependent variable** $Y$ is a random variable.
- The **independent variables** $X_1, \ldots, X_k$ are not, properly speaking, variables. Every independent variable $X$ is just a set of values $x_1, \ldots, x_n$. They can be predetermined values, like doses in a pharmaceutical study, a dummy coding gender or the year number, or observations of variables with a continuous distribution, with mean, variance, etc. In practical statistical analysis, we typically find a combination of these. In certain applications, which will be discussed in the Econometrics course, the values of the independent variables are assumed to have been obtained on a random sample of some population, so it makes sense to talk about the distribution, or about the correlation, of the independent variables. Anyway, the model does not include any distributional assumption for the $X$'s.
- The **error term** $\epsilon$ is a random variable with zero mean and variance $\sigma^2$. The assumption that the **error variance** $\sigma^2$ is the same for all the $X$ points is called **homoskedasticity** (sorry about this). When it is not valid, we say that there is **heteroskedasticity**. Methods for dealing with heteroskedasticity will be seen in the Econometrics course. $\epsilon$ is typically assumed to be normally distributed, but I postpone the discussion of this assumption to the next lecture. All the formulas presented in this lecture are valid without the normality assumption.
- The **regression coefficients** $\beta_0, \beta_1, \ldots, \beta_k$. The **intercept** or constant term, $\beta_0$, which could be seen as the expected value of $Y$ when $X_1 = \cdots = X_k = 0$, has rarely any interest, since that setting does not make sense in most applications. A **slope coefficient** $\beta_i$ is usually associated to the **effect**, or influence, of $X_i$ on $Y$ (see the discussion of Example 1).

It may seem to you that this definition is too broad, including too many possibilities for the independent variables. Indeed, it is. This is what gives linear regression its broad range of applications. The parameters of the model are the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and the error variance $\sigma^2$. This lecture deals with the estimation of these parameters.

## Ordinary least squares estimation

We have seen in the Mathematics course how to derive a linear equation from a data set. In mathematical terms, this can be described as an optimization problem: to minimize the sum of the squared residuals. Or, if you prefer, as an algebra problem: an orthogonal projection of the $Y$ vector on the subspace generated by the $X$ vectors. This is the least squares method, which, from now on, we will call **ordinary least squares** (OLS), to distinguish it from other variants of the same approach. When regarded as an estimation method, it is called **OLS estimation**.

Besides the assumptions about the error term discussed above, the observations (the rows) of the data set are assumed to be *statistically independent*. You may think that this remark is superfluous, since we have always made such assumption on the estimation and testing methods that we have seen in this course, but it is not so, since in econometric analysis you may have to deal with data for which this assumption is not realistic, for instance in longitudinal studies.

Packing the $Y$ values as an $n$-vector, the $X$ points as an $(n, k+1)$-matrix $\mathbf{X}$, the $\beta$'s as a $k$-vector coefficient $\boldsymbol{\beta}$ and the $\epsilon$ as a random $n$-vector $\boldsymbol{\epsilon}$,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

the regression equation becomes $\mathbf{y} = \mathbf{X}\beta + \epsilon$. The formula of the orthogonal projection gives us the OLS estimator of $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}.$$

Two satisfactory properties of the OLS estimator are unbiasedness and efficiency. Mind that, as in any other assertion about the distribution of the OLS estimators, it is assumed here that the matrix $\mathbf{X}$ is given. This is usually expressed as *conditional to* $\mathbf{X}$. For instance, we may write $\mathrm{E}\big[\mathbf{y}|\mathbf{X}\big] = \mathbf{X}\beta$, which is understood as giving the expected value of $Y$ for a given $X$ points as a linear combination of the values at that point.

To see that the OLS estimator is conditionally unbiased, we calculate the expectation,

$$\mathrm{E}\big[\hat{\boldsymbol{\beta}}|\mathbf{X}\big] = \mathrm{E}\big[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}|\mathbf{X}\big] = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\,\mathrm{E}\big[\mathbf{y}|\mathbf{X}\big] = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

For the efficiency, we calculate the covariance matrix of $\hat{\boldsymbol{\beta}}$. Again, everything is conditional to $\mathbf{X}$ but, now, I omit this detail in the notation. The assumptions of the indepedence of the observations and the homoskedasticity imply that $\mathrm{cov}[\boldsymbol{\epsilon}] = \mathbf{I}\sigma^2$, where $\mathbf{I}$ is the $n$ dimensional identity matrix. Therefore,

$$\mathrm{cov}\big[\hat{\boldsymbol{\beta}}\big] = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\,\mathrm{cov}[\mathbf{y}]\,\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}(\sigma^2\mathbf{I})\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\sigma^2.$$

It can be proved that this covariance matrix is minimal. The OLS estimator is the best linear estimator (BLUE) of the regression coefficients, meaning that it is more efficient than any other linear estimator. This property is called the **Gauss-Markov theorem**.

The estimate of the error variance $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2\,,$$

where the $e_i$ are the residuals. The square root $\hat{\sigma}$ is sometimes reported as the **residual standard error** (`R`) or the root MSE (Stata). The denominator $n-k-1$ is a bias correction.

## The regression line

I illustrate in this section the formulas of the preceding section with their application to the regression line. Now,

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

So,

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}.$$

It is easy to check that

$$\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} = \frac{1}{n\sum(x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}.$$

With a bit of algebra, we get

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \frac{1}{\sum(x_i - \bar{x})^2} \begin{bmatrix} \bar{y}\sum(x_i - \bar{x})^2 - \bar{x}\sum(x_i - \bar{x})(y_i - \bar{y}) \\ \sum(x_i - \bar{x})(y_i - \bar{y}) \end{bmatrix},$$

which is the same as the formulas of the intercept and the slope of the regression line. From the expression obtained for $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$ we can extract

$$\operatorname{var}[\hat{\beta}_0] = \frac{\sigma^2 \sum x_i^2}{n \sum(x_i - \bar{x})^2}, \qquad \operatorname{var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}, \qquad \operatorname{cov}[\hat{\beta}_0, \hat{\beta}_1] = \frac{-\bar{x}\,\sigma^2}{\sum(x_i - \bar{x})^2}.$$

Note that the estimators of the regression parameters are correlated (see exercise B). This implies that separate inferences about regression coefficients (this is what most people usually do) are wrong, since the coefficients must be taken together, as a whole.

The standard errors of $\hat{\beta}_0$ and $\beta_1$ are given by the square root of the variance. Since $\sigma^2$ is unknown, an estimate of the standard error is obtained replacing $\sigma^2$ by $\hat{\sigma}^2$:

$$\hat{\operatorname{se}}[\hat{\beta}_0] = \frac{\hat{\sigma} \sum x_i^2}{\sqrt{n \sum(x_i - \bar{x})^2}}, \qquad \hat{\operatorname{se}}[\hat{\beta}_1] = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}.$$

It is not hard to see that the standard errors get smaller as $n$ increases. As we will see in next lecture, this means, in practice, that everything becomes significant as the sample gets bigger. For the multiple regression case, the formulas are more involved, so matrix formulas are always used. The computer can take care of this.

**Example 1.** Modelling the labor market is an Econometrics classic. The Mroz87 data set, with data on 753 households, is used in many courses and textbooks. It is taken from the 1976 Panel Survey of Income Dynamics (PSID). The `mroz47.csv` file corresponds to a subsample formed by the 428 households where the wife has actually a job, and contains five variables:

- `wage`, 1975 wife's average earnings per hour, in US dollars.
- `wa`, wife's age, in years.
- `we`, wife's educational attainment, in years.
- `city`, a dummy for living in a large city.
- `exper`, wife's previous labor experience, in years.

I estimate first a linear regression model that relates the wife's wages to her educational attainment. The equation obtained ($R^2 = 0.117$), with the standard errors in parenthesis, is

$$\text{wage} = \underset{(0.848)}{-\,2.092} + \underset{(0.066)}{0.495}\ \text{we}.$$

Including the other three variables in the model, I get ($R^2 = 0.125$)

$$\text{wage} = -\underset{(1.224)}{2.794} + \underset{(0.022)}{0.007}\ \text{wa} + \underset{(0.067)}{0.483}\ \text{we} + \underset{(0.319)}{0.440}\ \text{city} + \underset{(0.021)}{0.021}\ \text{exper}.$$

This simple exercise illustrates several interesting facts:

- The second equation fits better the data, as shown by the increase of $R^2$. It has to be so, because the longer equation is optimal in a set of equations in which we can play with five coefficients, while in the shorter equation three of the coefficients are constrained to zero. Taking $R^2$ as a percentage of variance explained by the model, this change ($\Delta R^2$), does not look relevant. We will see later how to test this.
- Visual inspection suggests that the additional terms are not significantly different of zero. Indeed, the ratio of the coefficient estimate to the standard error is much smaller than the critical value $z = 1.96$ which defines the 95% confidence interval for the standard normal. This is consistent with the low value of $\Delta R^2$.
- As expected, the `we` coefficient changes when including additional terms in the equation, although the change is not relevant in this example. This coefficient has, in this case, a simple interpretation. Since it is positive, it indicates that the wages increase when the educational attainment increases. If we increase `we` by one year, `wa` is expected to increase, on the average, \$0.495 per hour. This is the usual interpretation of the slope: the average change of $Y$ when $X$ is increased by one unit. Nevertheless, this type of analysis has to be applied with care. First, it is valid only for feasible increments applied in the centre of the data set (in this case, `we` varies from 5 to 17 years). Second, it makes sense only in cross sectional data, in which the $X$ points can be taken as a statistical sample of a distribution.
- In the second equation, the interpretation of the coefficient is different. It is now the average change of $Y$ when increasing $X$ one unit, *holding the other independent variables constant*. This may seem a minor detail, but it is a capital fact, as we will see in the Econometrics course.

¶ Source: TA Mroz (1987), The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, *Econometrica* **55**, 765–799. □

### Homework

**A.** Because of the skewed distribution of variables such as salaries, sales or size, econometricians introduce them in log scale in linear models. Rerun the analysis of Example 1 replacing `wage` by `log(wage)`. How do you interpret the coefficients in both cases?

**B.** Generate 1,000 independent samples of size 10 of $(X, Y)$ as follows. Take $X$ and $\epsilon$ independent, with $X \sim \mathcal{N}(2, 1)$ and $\epsilon \sim \mathcal{N}(0, 0.04)$. Then define $Y$ as $Y = 3 + X + \epsilon$. For every sample, fit a regression line. Save the coefficients obtained and examine the joint distribution of the slope and the intercept.

Redo the exercise with $X \sim \mathcal{N}(0, 1)$. Explain the different results obtained.