[STAT-10] 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 discussing the different elements of the model and the assumptions involved. The elements are:

- The **dependent variable** Y is a random variable.
- Properly speaking, the **independent variables** $X_1, \ldots, X_k$ are not 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 or a dummy coding gender, but also observations of variables with a continuous distribution, with mean, variance, etc. In practical statistical analysis, we typically find a combination of these. In **crosss sectional data** analysis, 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** ϵ is a random variable with zero mean and variance σ^2. The assumption that the **error variance** σ^2 is the same for all the X points is called **homoskedasticity**. When it is not valid, we say that there is **heteroskedasticity**. Methods for dealing with heteroskedasticity will be discussed later in this course. The error term is typically assumed to be normally distributed, but I leave this assumption for later. 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, β_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** β_i is usually associated to the **effect**, or influence, of X_i on Y (see the discussion of Example 1).

This approach may look a bit loose, including too many possibilities for the independent variables. Indeed, it is, in order to give linear regression its broad range of applications.

Ordinary least squares estimation

The parameters of the linear model are the coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and the error variance σ^2. This lecture deals with estimating and testing these parameters. In particular, the estimation of the β's is based the OLS formulas. When regarded as an estimation method, we call it **OLS estimation**.

To get the desired properties for the OLS estimator, the error term observations (ϵ_i) are assumed to be *statistically independent*. You may think that this remark is superfluous, since we have always made such assumption so far, when discussing estimation and testing methods. It is not

so, since in econometric analysis we 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 β's as a k-vector coefficient $\boldsymbol{\beta}$ and the ϵ values 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 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 set of values of X_1, …, X_k, as a linear combination of these values.

Skipping the math detail, the (conditional) expectation and covariance of the OLS estimator are

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

So, the OLS estimator is conditionally unbiased. Also, it can be proved that this covariance matrix is minimal among the linear estimators. Hence, the OLS estimator is the best linear estimator (BLUE) of the regression coefficients. This property is called the **Gauss-Markov theorem**.

The estimate of the error variance σ^2 is

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

where the e_i's are the residuals. The square root $\hat{\sigma}$ is sometimes reported as the **residual standard error** (in R) or the root MSE (in 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,

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

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

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

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

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

Note that the estimators of the regression parameters are correlated. This implies that separate inferences about regression coefficients (which 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 β_1 are given by the square root of the variance. Since σ^2 is unknown, an estimate of the standard error is obtained by replacing σ^2 by $\hat{\sigma}^2$:

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

It is not hard to see that the standard errors get smaller as n increases. As we will see below, 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. Modeling 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

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

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

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

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 (ΔR^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 ΔR^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, `wage` 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, because it facilitates an interpretation in causal terms.

¶ 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.

The normality assumption

We see next how to test the coefficients of a linear regression equation. For testing, distributional assumptions are needed. The standard assumption is that the error term is normally distributed. We write this as $\epsilon \sim \mathcal{N}(0, \sigma^2)$. This looks quite straightforward, but a bit of confusion is typical around this point, so I include here some remarks:

- It is ϵ, not Y, what is assumed to be normal. This means that, for a fixed X point, Y has a normal distribution. For another point, Y will also be normal, but with a different mean. This is sometimes expressed saying that Y is conditionally normal. An example may help to clarify this. Take as Y the income in a certain population and as X a dummy for gender, and suppose that the average income is different in the two subpopulations. The distribution of Y, conditional to $X = 1$, is assumed to be normal, and the same for the distribution of Y conditional to $X = 0$, which means that the income is normally distributed in the two subpopulations. But when the two subpopulations are merged, the distribution is no longer normal. We will have a "camel" distribution here.
- Testing a coefficient usually means testing a null $H_0: \ \beta_i = 0$. When the null is rejected, we say that that coefficient is significant. Although a P-value is typically reported by stat packages for β_0, we practically never test the intercept, leaving it in the equation, significant or not.
- Many people believe that a nonsignificant coefficient means that the corresponding term *should* be dropped. It is not so. A better rule would be that it *could* be dropped.

What can be done without the normality assumption? Practically the same, for big samples. Indeed, it can be shown that the OLS estimators of the regression coefficients are asymptotically normally distributed, so that most of the previous discussion remains valid.

The t tests

Under the normality assumption, we can calculate $1 - 2\alpha$ confidence limits for the regression coefficients with the formula $\hat{\beta} \pm t_\alpha \, \hat{\text{se}}[\hat{\beta}]$. Standard errors can also be used to run a t test. For the null $H_0: \ \beta = 0$, we use

$$t = \frac{\hat{\beta}}{\hat{\text{se}}[\hat{\beta}]} \, ,$$

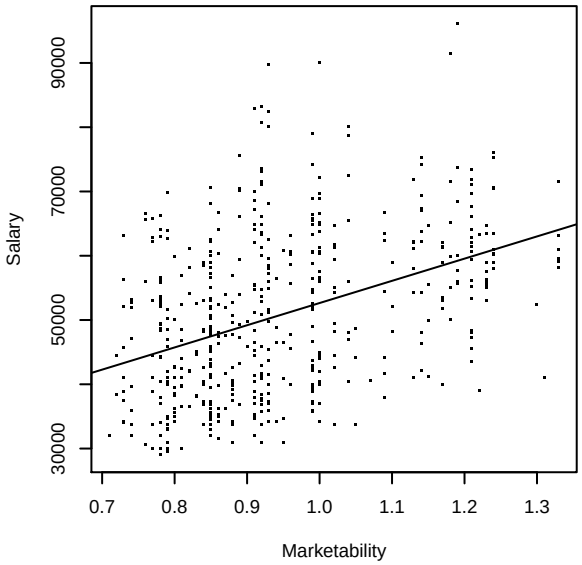

Figure 1. Regression line (Example 2)

with $\mathrm{df} = n - k - 1$. Note that the standard error is not the exact value, but an estimate. That is why we use the Student t instead of the standard normal. In the computer, the standard regression output contains, besides every parameter estimate, the standard error, the t statistic and the P-value (Table 1). Also, note that, for very big samples, t statistics are likely to be significant.

The F test

In many stat packages, the default regression report includes the ANOVA decomposition (it is so in Stata, but not in `R`). The F statistic associated to this ANOVA decomposition,

$$F = \big(n-2\big)\,\frac{\mathrm{SSE}}{\mathrm{SSR}} = \frac{(n-2)\,R^2}{1-R^2}$$

is used to test the null $H_0:\ \beta_1 = \cdots = \beta_k = 0$.

The second expression of this statistic, involving R^2, shows that significance occurs when R^2 is close enough to 1. We say then that R^2 is significant, or that the (multiple) correlation is significant. It is also obvious, from the formula, that weak correlations are significant with samples big enough. In the simple regression case, the F statistic is the square of the t statistic associated to β_1, so it is redundant (this is no longer true in multiple regression).

Residual analysis

The analysis of the residuals is useful for checking the validity of the model. This analysis is similar to that of the residuals of the one-way ANOVA, plus the possibility of a **residual plot**. In a residual plot, we place the residuals (standardized or not) in the ordinates, and either one independent variable, the predicted values, or the order in which the data were obtained, in the abscissas.

Example 2. The `market` data set contains data from a study on the salaries of academic staff in Bowling Green State University. This data set has been used in several textbooks and can be considered as a standard example. The sample size is 514. We use here the variables:

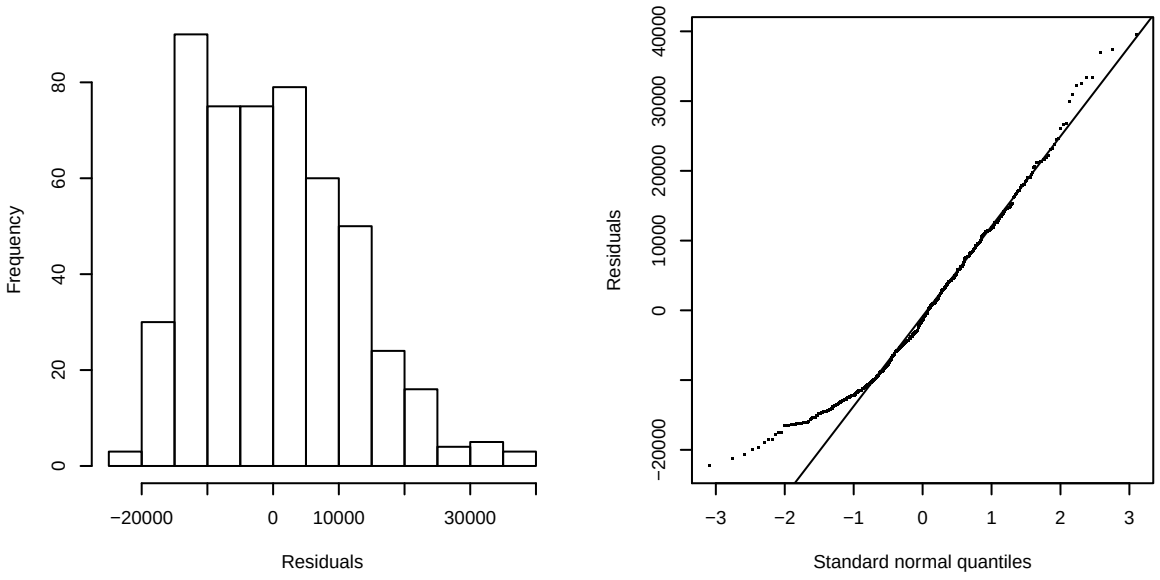

Figure 2. Distribution of residuals (Example 2)

- `salary`, academic year (9 month) salary in US dollars.
- `market`, marketability of the discipline, defined as the ratio of national average salary paid in the discipline to the national average across disciplines.

It is natural here to take marketability as the independent variable, fitting a regression line to the data to produce an equation for predicting the salary from the marketability. The R-squared statistic is $R^2 = 0.166$. In the coefficients table (Table 1), we find the coefficient estimates, the standard errors, the t statistic and the p-value.

TABLE 1. Linear regression results (Example 1)

Coefficient	Estimate	Std. error	t statistic	p-value
Intercept	18,097.0	3,288.0	5.50	0.000
`market`	34,545.2	3,424.3	10.09	0.000

The F statistic is $F(1, 512) = 101.8$. Note that this is the same as the square of the t statistic for marketability ($t = 10.09$), with the same p-value.

The diagnostic plots of Figure 2 reveal a clear departure from normality on the left tail. Indeed, the skewness is Sk $= 0.595$, the kurtosis K $= -0.017$, and the Jarque-Bera statistic JB $= 30.3$ ($P < 0.001$). Nevertheless, given the sample size, non-normality is not a problem here.

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?