Regression Analysis: Basic Concepts

Allin Cottrell

1 The simple linear model

Suppose we reckon that some variable of interest, y, is 'driven by' some other variable x. We then call y the *dependent* variable and x the *independent* variable. In addition, suppose that the relationship between y and x is basically linear, but is inexact: besides its determination by x, y has a random component, u, which we call the 'disturbance' or 'error'.

Let i index the observations on the data pairs (x, y). The simple linear model formalizes the ideas just stated:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

The parameters β_0 and β_1 represent the y-intercept and the slope of the relationship, respectively.

In order to work with this model we need to make some assumptions about the behavior of the error term. For now we'll assume three things:

$$
\begin{array}{ll}
E(u_i) = 0 & u \text{ has a mean of zero for all } i \\
E(u_i^2) = \sigma_u^2 & \text{it has the same variance for all } i \\
E(u_i u_j) = 0, i \neq j & \text{no correlation across observations}
\end{array}
$$

We'll see later how to check whether these assumptions are met, and also what resources we have for dealing with a situation where they're not met.

We have just made a bunch of assumptions about what is 'really going on' between y and x, but we'd like to put numbers on the parameters β_0 and β_1. Well, suppose we're able to gather a sample of data on x and y. The task of *estimation* is then to come up with coefficients—numbers that we can calculate from the data, call them $\hat{\beta}_0$ and $\hat{\beta}_1$—which serve as estimates of the unknown parameters.

If we can do this somehow, the estimated equation will have the form

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x.$$

We define the estimated error or *residual* associated with each pair of data values as the actual y_i value minus the prediction based on x_i along with the estimated coefficients

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right)$$

In a scatter diagram of y against x, this is the vertical distance between observed y_i and the 'fitted value', $\hat{y}_i$, as shown in Figure 1.

Note that we are using a different symbol for this *estimated* error ($\hat{u}_i$) as opposed to the 'true' disturbance or error term defined above (u_i). These two will coincide only if $\hat{\beta}_0$ and $\hat{\beta}_1$ happen to be exact estimates of the regression parameters β_0 and β_1.

The most common technique for determining the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ is Ordinary Least Squares (OLS): values for $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen so as to minimize the sum of the squared residuals or SSR. The SSR may be written as

$$\text{SSR} = \Sigma \hat{u}_i^2 = \Sigma (y_i - \hat{y}_i)^2 = \Sigma (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

(It should be understood throughout that Σ denotes the summation $\sum_{i=1}^{n}$, where n is the number of observations in the sample). The minimization of SSR is a calculus exercise: we need to find the partial derivatives of SSR
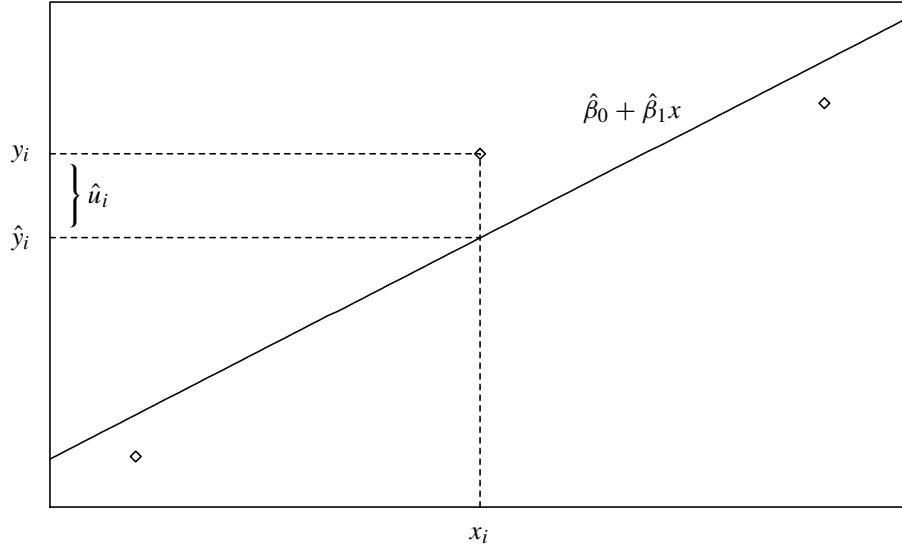
Figure 1: Regression residual

with respect to both $\hat{\beta}_0$ and $\hat{\beta}_1$ and set them equal to zero. This generates two equations (known as the 'normal equations' of least squares) in the two unknowns, $\hat{\beta}_0$ and $\hat{\beta}_1$. These equations are then solved jointly to yield the estimated coefficients.

We start out from:

$$\partial \text{SSR}/\partial \hat{\beta}_0 = -2\Sigma(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{1}$$

$$\partial \text{SSR}/\partial \hat{\beta}_1 = -2\Sigma x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{2}$$

Equation (1) implies that

$$\Sigma y_i - n\hat{\beta}_0 - \hat{\beta}_1 \Sigma x_i = 0$$

$$\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3}$$

while equation (2) implies that

$$\Sigma x_i y_i - \hat{\beta}_0 \Sigma x_i - \hat{\beta}_1 \Sigma x_i^2 = 0 \tag{4}$$

We can now substitute for $\hat{\beta}_0$ in equation (4), using (3). This yields

$$\Sigma x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x})\Sigma x_i - \hat{\beta}_1 \Sigma x_i^2 = 0$$

$$\Rightarrow \Sigma x_i y_i - \bar{y}\Sigma x_i - \hat{\beta}_1(\Sigma x_i^2 - \bar{x}\Sigma x_i) = 0$$

$$\Rightarrow \hat{\beta}_1 = \frac{\Sigma x_i y_i - \bar{y}\Sigma x_i}{\Sigma x_i^2 - \bar{x}\Sigma x_i} \tag{5}$$

Equations (3) and (4) can now be used to generate the regression coefficients. First use (5) to find $\hat{\beta}_1$, then use (3) to find $\hat{\beta}_0$.

2 Goodness of fit

The OLS technique ensures that we find the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ which 'fit the sample data best', in the specific sense of minimizing the sum of squared residuals. There is no guarantee, however, that $\hat{\beta}_0$ and $\hat{\beta}_1$ correspond

exactly with the unknown parameters β_0 and β_1. Neither, in fact, is there any guarantee that the 'best fitting' line fits the data well: maybe the data do not even approximately lie along a straight line relationship. So how do we assess the adequacy of the 'fitted' equation?

First step: find the residuals. For each x-value in the sample, compute the fitted value or predicted value of y, using $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Then subtract each fitted value from the corresponding actual, observed, value of y_i. Squaring and summing these differences gives the SSR, as shown in Table 1. In this example, based on a sample of 14 houses, y_i is sale price in thousands of dollars and x_i is square footage of living area.

Table 1: Example of finding residuals

Given $\hat{\beta}_0 = 52.3509$; $\hat{\beta}_1 = 0.1388$

data (x_i)	data (y_i)	fitted ($\hat{y}_i$)	$\hat{u}_i = y_i - \hat{y}_i$	$\hat{u}_i^2$
1065	199.9	200.1	−0.2	0.04
1254	228.0	226.3	1.7	2.89
1300	235.0	232.7	2.3	5.29
1577	285.0	271.2	13.8	190.44
1600	239.0	274.4	−35.4	1253.16
1750	293.0	295.2	−2.2	4.84
1800	285.0	302.1	−17.1	292.41
1870	365.0	311.8	53.2	2830.24
1935	295.0	320.8	−25.8	665.64
1948	290.0	322.6	−32.6	1062.76
2254	385.0	365.1	19.9	396.01
2600	505.0	413.1	91.9	8445.61
2800	425.0	440.9	−15.9	252.81
3000	415.0	468.6	−53.6	2872.96
			$\Sigma = 0$	$\Sigma = 18273.6$
				= SSR

Now, obviously, the magnitude of the SSR will depend in part on the number of data points in the sample (other things equal, the more data points, the bigger the sum of squared residuals). To allow for this we can divide though by the 'degrees of freedom', which is the number of data points minus the number of parameters to be estimated (2 in the case of a simple regression with an intercept term). Let n denote the number of data points (or 'sample size'), then the degrees of freedom, d.f. $= n - 2$. The square root of the resulting expression is called the estimated *standard error* of the regression ($\hat{\sigma}$):

$$\hat{\sigma} = \sqrt{\frac{\text{SSR}}{n-2}}$$

The standard error gives us a first handle on how well the fitted equation fits the sample data. But what is a 'big' $\hat{\sigma}$ and what is a 'small' one depends on the context. The standard error is sensitive to the units of measurement of the dependent variable.

A more standardized statistic, which also gives a measure of the 'goodness of fit' of the estimated equation, is R^2. This statistic (sometimes known as the coefficient of determination) is calculated as follows:

$$R^2 = 1 - \frac{\text{SSR}}{\Sigma(y_i - \bar{y})^2} \equiv 1 - \frac{\text{SSR}}{\text{SST}}$$

Note that SSR can be thought of as the 'unexplained' variation in the dependent variable—the variation 'left over' once the predictions of the regression equation are taken into account. The expression $\Sigma(y_i - \bar{y})^2$, on the other hand, represents the *total variation* (total sum of squares or SST) of the dependent variable around its mean value. So R^2 can be written as 1 minus the proportion of the variation in y_i that is 'unexplained'; or in other words it

shows *the proportion of the variation in y_i that is accounted for by the estimated equation*. As such, it must be bounded by 0 and 1.

$$0 \leq R^2 \leq 1$$

$R^2 = 1$ is a 'perfect score', obtained only if the data points happen to lie exactly along a straight line; $R^2 = 0$ is perfectly lousy score, indicating that x_i is absolutely useless as a predictor for y_i.

To summarize: alongside the estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, we can also examine the sum of squared residuals (SSR), the regression standard error ($\hat{\sigma}$) and/or the R^2 value, in order to judge whether the best-fitting line does in fact fit the data to an adequate degree.