

Regression Analysis: Basic Concepts

Allin Cottrell*

1 The simple linear model

This model represents the dependent variable, y_i , as a linear function of one independent variable, x_i , subject to a random ‘disturbance’ or ‘error’, u_i .

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

The error term u_i is assumed to have a mean value of zero, a constant variance, and to be uncorrelated with itself across observations ($E(u_i u_j) = 0, i \neq j$). We may summarize these conditions by saying that u_i ‘white noise’. The task of estimation is to determine regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, estimates of the unknown parameters β_0 and β_1 respectively. The estimated equation will have the form

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

We define the estimated error or *residual* associated with each pair of data values as

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

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

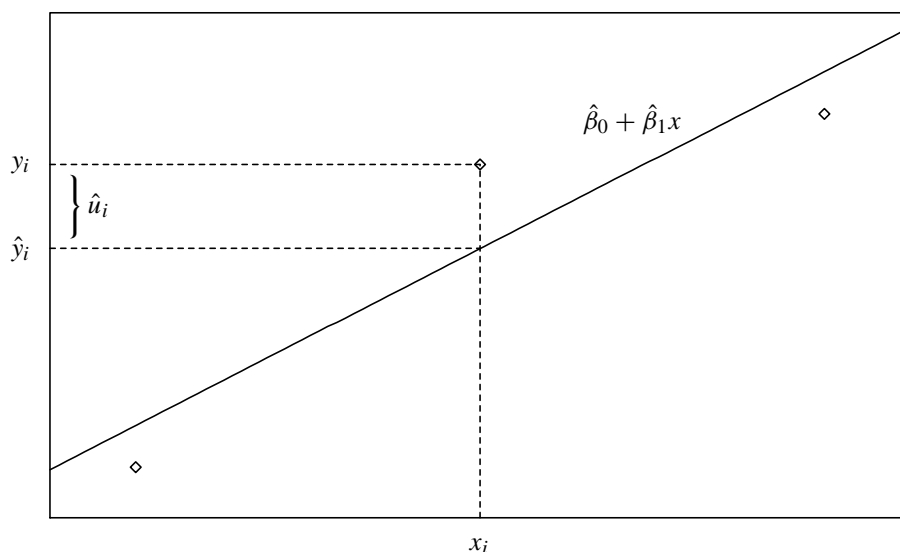


Figure 1: Regression residual

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 basic 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 (SSR). The SSR may be written as

*Last revised 2003/02/03.

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

(It should be understood throughout that Σ denotes the summation $\sum_{i=1}^n$, where n denotes the number of observations in the sample). The minimization of SSR is a calculus exercise: we need to find the partial derivatives of SSR with respect to both $\hat{\beta}_0$ and $\hat{\beta}_1$ and set them equal to zero. This generates two equations (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:

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

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

Equation (1) implies that

$$\begin{aligned} \Sigma y_i - n \hat{\beta}_0 - \hat{\beta}_1 \Sigma x_i &= 0 \\ \Rightarrow \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \quad (3)$$

while equation (2) implies that

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

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

$$\begin{aligned} \Sigma x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \Sigma x_i - \hat{\beta}_1 \Sigma x_i^2 &= 0 \\ \Rightarrow \Sigma x_i y_i - \bar{y} \Sigma x_i - \hat{\beta}_1 (\Sigma x_i^2 - \bar{x} \Sigma x_i) &= 0 \\ \Rightarrow \hat{\beta}_1 &= \frac{\Sigma x_i y_i - \bar{y} \Sigma x_i}{\Sigma x_i^2 - \bar{x} \Sigma x_i} \end{aligned} \quad (5)$$

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.

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 ($\hat{\sigma}$):

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

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

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 is calculated as follows:

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

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 .

When you add an additional variable to a regression equation, there is no way it can raise the SSR, and in fact it’s likely to lower the SSR somewhat even if the added variable is not very relevant. And lowering the SSR means raising the R^2 value. One might therefore be tempted to add too many extraneous variables to a regression if one were focussed on achieving the maximum R^2 . An alternative calculation, the adjusted R-squared or \bar{R}^2 , attaches a small penalty to adding more variables: thus if adding an additional variable raises the \bar{R}^2 for a regression, that’s a better indication that it has “improved” the model than if it merely raises the plain, unadjusted R^2 . The formula is:

$$\bar{R}^2 = 1 - \frac{SSR/(n - k - 1)}{SST/(n - 1)} = 1 - \frac{n - 1}{n - k - 1}(1 - R^2)$$

where $k + 1$ represents the number of parameters being estimated (2 in a simple regression).

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

3 Confidence intervals for regression coefficients

As stated above, even if the OLS math is performed correctly there is no guarantee that the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ thus obtained correspond exactly with the underlying parameters β_0 and β_1 . Actually, such an exact correspondence is highly unlikely. The statistical issue here is a very general one: *estimation is inevitably subject to sampling error.*

As we have seen, a *confidence interval* provides a means of quantifying the uncertainty produced by sampling error. Instead of simply stating ‘I found a sample mean income of \$39,000 and that is my best guess at the population mean, although I know it is probably wrong’, we can make a statement like: ‘I found a sample mean of \$39,000, and there is a 95 percent probability that my estimate is off the true parameter value by no more than \$1200.’

Confidence intervals for regression coefficients can be constructed in a similar manner. Suppose we come up with a slope estimate of $\hat{\beta}_1 = .90$, using the OLS technique, and we want to quantify our uncertainty over the true slope parameter, β_1 , by drawing up a 95 percent confidence interval for this parameter.

Provided our sample size is reasonably large, the rule of thumb is the same as before; the 95 percent confidence interval for β_1 is given by:

$$\hat{\beta}_1 \pm 2 \text{ standard errors}$$

Our single best guess at β_1 (‘point estimate’) is simply $\hat{\beta}_1$, since the OLS technique yields unbiased estimates of the parameters (actually, this is not *always* true, but we’ll postpone consideration of tricky cases where OLS estimates are biased). And on exactly the same grounds as before, there is a 95 per chance that our estimate $\hat{\beta}_1$ will lie within 2 standard errors of its mean value, β_1 . But how do we find the standard error of $\hat{\beta}_1$? I shall not derive this rigorously, but give the formula along with an intuitive explanation. The standard error of $\hat{\beta}_1$ (written as $se(\hat{\beta}_1)$, and not to be confused with the standard error of the regression, $\hat{\sigma}$) is given by the formula:

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2}}$$

i.e., it is the square root of [the square of the regression standard error divided by the total variation of the independent variable, x_i , around its mean].

What are the various components of the calculation doing? First, note the general point that the larger is $se(\hat{\beta}_1)$, the wider will be the confidence interval for any specified confidence level. Now, according to the formula, the larger is $\hat{\sigma}$, the larger will be $se(\hat{\beta}_1)$, and hence the wider the confidence interval for the true slope. This makes sense: $\hat{\sigma}$ provides a measure of the ‘degree of fit’ of the estimated equation, as discussed above. If the equation fits the data badly (‘large’ $\hat{\sigma}$), it stands to reason that we should have a relatively high degree of uncertainty over the true slope parameter.

Secondly, the formula tells us that, other things equal, a high degree of variation of x_i makes for a smaller $se(\hat{\beta}_1)$, and so a tighter confidence interval. Why should this be? The more x_i has varied in our data sample, the better the chance we have of picking up any relationship that exists between x and y . Take an extreme case and this is rather obvious: suppose that x happens not to have varied at all in our sample (i.e., $\sum(x_i - \bar{x})^2 = 0$). In that case we have no chance of detecting any influence of x on y . And the more the independent variable has moved, the more any influence it may have on the dependent variable should stand out against the background ‘noise’, u_i .

4 Example of confidence interval for the slope parameter

One example. Suppose we’re interested in whether a positive linear relationship exists between x_i and y_i . We’ve obtained $\hat{\beta}_1 = .90$ and $se(\hat{\beta}_1) = .12$. The approximate 95 percent confidence interval for β_1 is then

$$.90 \pm 2(.12) = .90 \pm .24 = .66 \text{ to } 1.14$$

This tells us that we can state, with at least 95 percent confidence, that $\beta_1 > 0$, and there is a positive relationship. On the other hand, if we had obtained $se(\hat{\beta}_1) = .61$, our interval would have been

$$.90 \pm 2(.61) = .90 \pm 1.22 = -.32 \text{ to } 2.12$$

In this case the interval straddles zero, and we cannot be confident (at the 95 percent level) that there exists a positive relationship.