Regression Analysis: Basic Concepts

Allin Cottrell

The simple linear 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, and to be uncorrelated with the independent variable, x. In the simplest case it is also assumed to have a constant variance, and to be uncorrelated with its own past values (i.e., it is "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

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

OLS

The basic technique for determining the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ is *Ordinary Least Squares* (OLS).

Values for the coefficients are chosen to minimize the sum of the squared estimated errors or *residual sum of squares* (SSR). The estimated error associated with each pair of data-values (x_i, y_i) is defined as

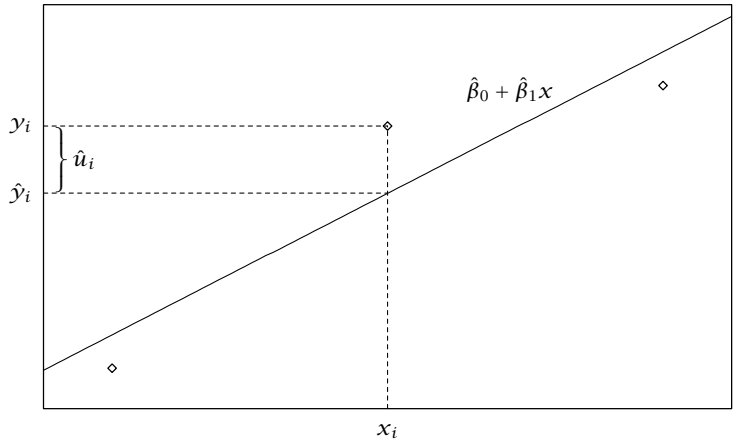
$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

We use a different symbol for this *estimated* error $(\hat{u}_i)$ as opposed to the "true" disturbance or error term, (u_i). These two coincide only if $\hat{\beta}_0$ and $\hat{\beta}_1$ happen to be exact estimates of the regression parameters α and β. The estimated errors are also known as *residuals*.

The SSR may be written as

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

Picturing the residuals



The residual, $\hat{u}_i$, is the vertical distance between the actual value of the dependent variable, y_i, and the fitted value, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Normal Equations

Minimization of SSR is a calculus exercise: 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 solved jointly to yield the estimated coefficients.

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

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

Equation (1) implies that

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

$$\Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \quad (3)$$

Equation (2) implies that

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

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

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

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

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

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 $\hat{\beta}_0$.

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 sense of minimizing the sum of squared residuals.

There's no guarantee that $\hat{\beta}_0$ and $\hat{\beta}_1$ correspond exactly with the unknown parameters β_0 and β_1. No guarantee that the "best fitting" line fits the data well at all: maybe the data do not even approximately lie along a straight line relationship. 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.

Example of finding residuals

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

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

Standard error

The magnitude of SSR depends in part on the number of data points. 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 intercept).

Let n denote the number of data points (sample size); then the degrees of freedom, df $= n - 2$.

The square root of (SSR/df) is the *standard error of the regression*, $\hat{\sigma}$:

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

The standard error gives 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 regression standard error is sensitive to the units of measurement of the dependent variable.

R-squared

A more standardized statistic which also gives a measure of the goodness of fit of the estimated equation is R^2.

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

- 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.
- $\Sigma(y_i - \bar{y})^2$ (total sum of squares or SST), represents the *total variation* of the dependent variable around its mean value.

$R^2 = (1 - \text{SSR/SST})$ is 1 minus the proportion of the variation in y_i that is unexplained.

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.

Adjusted R-squared

Adding a variable to a regression equation cannot raise the SSR; it's likely to lower SSR somewhat even if the new variable is not very relevant.

The *adjusted* R-squared, $\bar{R}^2$, attaches a small penalty to adding more variables. If adding a variable raises the $\bar{R}^2$ for a regression, that's a better indication that is has improved the model that if it merely raises the unadjusted R^2.

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

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 might also examine

- the sum of squared residuals (SSR)
- the regression standard error ($\hat{\sigma}$)
- the R^2 value (adjusted or unadjusted)

to judge whether the best-fitting line does in fact fit the data to an adequate degree.

Confidence intervals for coefficients

As we saw, 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 are constructed in a similar manner.

Suppose we're interested in the slope coefficient, $\hat{\beta}_1$, of an estimated equation. Say we came up with $\hat{\beta}_1 = .90$, 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 β_1.

Provided our sample size is reasonably large, the rule of thumb is the same as before; the 95 percent confidence interval for β 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).

On 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.

The standard error of $\hat{\beta}_1$ (written as $\text{se}(\hat{\beta}_1)$, and not to be confused with the standard error of the regression, $\hat{\sigma}$) is given by the formula:

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

- The larger is $\text{se}(\hat{\beta}_1)$, the wider will be our confidence interval.
- The larger is $\hat{\sigma}$, the larger will be $\text{se}(\hat{\beta}_1)$, and so the wider the confidence interval for the true slope. Makes sense: in the case of a poor fit we have high uncertainty over the true slope parameter.
- A high degree of variation of x_i makes for a smaller $\text{se}(\hat{\beta}_1)$ (tighter confidence interval). The more x_i has varied in our sample, the better the chance we have of accurately picking up any relationship that exists between x and y.

Confidence interval example

Is there really a positive linear relationship between x_i and y_i? We've obtained $\hat{\beta}_1 = .90$ and $\text{se}(\hat{\beta}_1) = .12$. The approximate 95 percent confidence interval for β_1 is then

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

Thus we can state, with at least 95 percent confidence, that $\beta_1 > 0$, and there is a positive relationship.

If we had obtained $\text{se}(\hat{\beta}_1) = .61$, 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. □