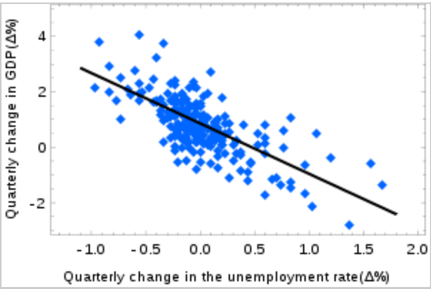
Simple linear regression

From Wikipedia, the free encyclopedia

In statistics, **simple linear regression** is the least squares estimator of a linear regression model with a single explanatory variable. In other words, simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared *residuals* of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible.



Okun's law in macroeconomics is an example of the simple linear regression. Here the dependent variable (GDP growth) is presumed to be in a linear relationship with the changes in the unemployment rate.

The adjective *simple* refers to the fact that this regression is one of the simplest in statistics. The slope of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that it passes through the center of mass (x, y) of the data points.

Other regression methods besides the simple ordinary least squares (OLS) also exist (see linear regression model). In particular, when one wants to do regression by eye, people usually tend to draw a slightly steeper line, closer to the one produced by the total least squares method. This occurs because it is more natural for one's mind to consider the orthogonal distances from the observations to the regression line, rather than the vertical ones as OLS method does.

Contents

Fitting the regression line

Suppose there are n data points $\{(x_i, y_i), i = 1, ..., n\}$. The goal is to find the equation of the straight line

$$y = \alpha + \beta x,$$

which would provide a "best" fit for the data points. Here the "best" will be understood as in the least-squares approach: a line that minimizes the sum of squared residuals of the linear regression model. In other words, α (the y-intercept) and β (the slope) solve the following minimization problem:

$$\text{Find } \min_{\alpha,\,\beta} Q(\alpha, \beta), \qquad \text{for } Q(\alpha, \beta) = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

By using either calculus, the geometry of inner product spaces or simply expanding to get a quadratic in α and β, it can be shown that the values of α and β that minimize the objective function Q[1] are

$$\begin{aligned}\hat\beta &= \frac{\sum_{i=1}^{n}(x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^{n}(x_i-\bar x)^2}\\ &= \frac{\sum_{i=1}^{n}x_iy_i - \frac{1}{n}\sum_{i=1}^{n}x_i\sum_{j=1}^{n}y_j}{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2}\\ &= \frac{\overline{xy}-\bar x\bar y}{\overline{x^2}-\bar x^2}\\ &= \frac{\operatorname{Cov}[x,y]}{\operatorname{Var}[x]}\\ &= r_{xy}\frac{s_y}{s_x},\end{aligned}$$

$$\hat\alpha = \bar y - \hat\beta\,\bar x,$$

where r_{xy} is the sample correlation coefficient between x and y; s_x is the standard deviation of x; and s_y is correspondingly the standard deviation of y. A horizontal bar over a quantity indicates the sample-average of that quantity. For example:

$$\overline{xy} = \tfrac{1}{n}\sum_{i=1}^{n}x_iy_i.$$

Substituting the above expressions for $\hat\alpha$ and $\hat\beta$ into

$$y = \hat\alpha + \hat\beta x,$$

yields

$$\frac{y-\bar y}{s_y} = r_{xy}\frac{x-\bar x}{s_x}$$

This shows the role r_{xy} plays in the regression line of standardized data points. It is sometimes useful to calculate r_{xy} from the data independently using this equation:

$$r_{xy} = \frac{\overline{xy}-\bar x\bar y}{\sqrt{(\overline{x^2}-\bar x^2)(\overline{y^2}-\bar y^2)}}$$

The coefficient of determination (R squared) is equal to r_{xy}^2 when the model is linear with a single independent variable. See sample correlation coefficient for additional details.

Linear regression without the intercept term

Sometimes, people consider a simple linear regression model without the intercept term, $y = \beta x$. In such a case, the OLS estimator for β simplifies to

$$\hat\beta = \frac{\overline{xy}}{\overline{x^2}}$$

and the sample correlation coefficient becomes

$$r_{xy} = \frac{\overline{xy}}{\sqrt{(\overline{x^2})(\overline{y^2})}}$$

Numerical properties

1. The line goes through the "center of mass" point $(\bar x, \bar y)$.
2. The sum of the residuals is equal to zero, if the model includes a constant: $\sum_{i=1}^{n}\hat\varepsilon_i = 0.$

3. The linear combination of the residuals, in which the coefficients are the x-values, is equal to zero: $\sum_{i=1}^n x_i\hat{\varepsilon}_i = 0.$

Model-cased properties

Description of the statistical properties of estimators from the simple linear regession estimates requires the use of a statistical model. The following is based on assuming the validity of a model under which the estimates are optimal. It is also possible to evaluate the properties under other assumptions, such as inhomogeneity, but this is discussed elsewhere.

Unbiasedness

The estimators $\hat{\alpha}$ and $\hat{\beta}$ are unbiased. This requires that we interpret the estimators as random variables and so we have to assume that, for each value of x, the corresponding value of y is generated as a mean response $\alpha + \beta x$ plus an additional random variable ε called the *error term*. This error term has to be equal to zero on average, for each value of x. Under such interpretation, the least-squares estimators $\hat{\alpha}$ and $\hat{\beta}$ will themselves be random variables, and they will unbiasedly estimate the "true values" α and β.

Confidence intervals

The formulas given in the previous section allow one to calculate the *point estimates* of α and β — that is, the coefficients of the regression line for the given set of data. However, those formulas don't tell us how precise the estimates are, i.e., how much the estimators $\hat{\alpha}$ and $\hat{\beta}$ vary from sample to sample for the specified sample size. So-called *confidence intervals* were devised to give a plausible set of values the estimates might have if one repeated the experiment a very large number of times.

The standard method of constructing confidence intervals for linear regression coefficients relies on the normality assumption, which is justified if either:

1. the errors in the regression are normally distributed (the so-called *classic regression* assumption), or
2. the number of observations n is sufficiently large, in which case the estimator is approximately normally distributed.

The latter case is justified by the central limit theorem.

Normality assumption

Under the first assumption above, that of the normality of the error terms, the estimator of the slope coefficient will itself be normally distributed with mean β and variance $\sigma^2/\sum(x_i-\bar{x})^2$, where σ^2 is the variance of the error terms (see Proofs involving ordinary least squares). At the same time the sum of squared residuals Q is distributed proportionally to χ^2 with $n-2$ degrees of freedom, and independently from $\hat{\beta}$. This allows us to construct a t-statistic

$$t = \frac{\hat{\beta}-\beta}{s_{\hat{\beta}}}\ \sim\ t_{n-2},$$

where

$$s_{\hat{\beta}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^n \hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^n (x_i-\bar{x})^2}}$$

which has a Student's t-distribution with $n-2$ degrees of freedom. Here s_β is the *standard error* of the estimator $\hat{\beta}$.

Using this t-statistic we can construct a confidence interval for β:

$$\beta \in \left[\hat{\beta} - s_{\hat{\beta}} t^*_{n-2},\ \hat{\beta} + s_{\hat{\beta}} t^*_{n-2}\right],$$

at confidence level $(1-\gamma)$, where t^*_{n-2} is the $(1-\frac{\gamma}{2})$-th quantile of the t_{n-2} distribution. For example, if $\gamma = 0.05$ then the confidence level is 95%.

Similarly, the confidence interval for the intercept coefficient α is given by

$$\alpha \in \left[\hat\alpha - s_{\hat\alpha} t^*_{n-2},\ \hat\alpha + s_{\hat\alpha} t^*_{n-2}\right],$$

at confidence level $(1-\gamma)$, where

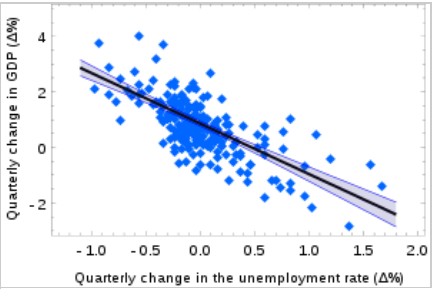
$$s_{\hat\alpha} = s_{\hat\beta}\sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^n x_i^2} = \sqrt{\tfrac{1}{n(n-2)}\left(\textstyle\sum_{j=1}^n \hat\varepsilon_j^2\right)\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i-\bar x)^2}}$$

The confidence intervals for α and β give us the general idea where these regression coefficients are most likely to be. For example in the "Okun's law" regression shown at the beginning of the article the point estimates are

$$\hat\alpha = 0.859, \qquad \hat\beta = -1.817.$$

The 95% confidence intervals for these estimates are

$$\alpha \in [0.76, 0.96], \qquad \beta \in [-2.06, -1.58].$$



The US "changes in unemployment – GDP growth" regression with the 95% confidence bands.

In order to represent this information graphically, in the form of the confidence bands around the regression line, one has to proceed carefully and account for the joint distribution of the estimators. It can be shown that at confidence level $(1-\gamma)$ the confidence band has hyperbolic form given by the equation

$$\hat y|_{x=\xi} \in \left[\hat\alpha + \hat\beta\xi \pm t^*_{n-2}\sqrt{\frac{1}{n-2}\sum\hat\varepsilon_i^2\cdot\left(\frac{1}{n} + \frac{(\xi-\bar x)^2}{\sum(x_i-\bar x)^2}\right)}\right].$$

Asymptotic assumption

The alternative second assumption states that when the number of points in the dataset is "large enough", the law of large numbers and the central limit theorem become applicable, and then the distribution of the estimators is approximately normal. Under this assumption all formulas derived in the previous section remain valid, with the only exception that the quantile t^*_{n-2} of Student's t distribution is replaced with the quantile q^* of the standard normal distribution. Occasionally the fraction $\frac{1}{n-2}$ is replaced with $\frac{1}{n}$. When n is large such change does not alter the results appreciably.

Numerical example

This example concerns the data set from the Ordinary least squares article. This data set gives average weights for humans as a function of their height in the population of American women of age 30–39. Although the OLS article argues that it would be more appropriate to run a quadratic regression for this data, the simple linear regression model is applied here instead.

x_i	1.47	1.50	1.52	1.55	1.57	1.60	1.63	1.65	1.68	1.70	1.73	1.75	1.78	1.80	1.83	**Height (m)**
y_i	52.21	53.12	54.48	55.84	57.20	58.57	59.93	61.29	63.11	64.47	66.28	68.10	69.92	72.19	74.46	**Mass (kg)**

There are n = 15 points in this data set. Hand calculations would be started by finding the following five sums:

$$S_x = \sum x_i = 24.76, \quad S_y = \sum y_i = 931.17$$
$$S_{xx} = \sum x_i^2 = 41.0532, \quad S_{xy} = \sum x_i y_i = 1548.2453, \quad S_{yy} = \sum y_i^2 = 58498.5439$$

These quantities would be used to calculate the estimates of the regression coefficients, and their standard errors.

$$\hat\beta = \frac{nS_{xy} - S_x S_y}{nS_{xx} - S_x^2} = 61.272$$
$$\hat\alpha = \tfrac{1}{n}S_y - \hat\beta \tfrac{1}{n}S_x = -39.062$$
$$s_\varepsilon^2 = \tfrac{1}{n(n-2)}\left(nS_{yy} - S_y^2 - \hat\beta^2(nS_{xx} - S_x^2)\right) = 0.5762$$
$$s_\beta^2 = \frac{n s_\varepsilon^2}{nS_{xx} - S_x^2} = 3.1539$$
$$s_\alpha^2 = s_\beta^2 \tfrac{1}{n} S_{xx} = 8.63185$$

The 0.975 quantile of Student's *t*-distribution with 13 degrees of freedom is t^*_{13} = 2.1604, and thus the 95% confidence intervals for α and β are

$$\alpha \in [\,\hat\alpha \mp t^*_{13} s_\alpha\,] = [\,-45.4,\ -32.7\,]$$
$$\beta \in [\,\hat\beta \mp t^*_{13} s_\beta\,] = [\,57.4,\ 65.1\,]$$

The product-moment correlation coefficient might also be calculated:

$$\hat r = \frac{nS_{xy} - S_x S_y}{\sqrt{(nS_{xx} - S_x^2)(nS_{yy} - S_y^2)}} = 0.9945$$

This example also demonstrates that sophisticated calculations will not overcome the use of badly prepared data. The heights were originally given in inches, and have been converted to the nearest centimetre. Since the conversion factor is one inch to 2.54 cm, this is *not* a correct conversion. The original inches can be recovered by Round(x/0.0254) and then re-converted to metric: if this is done, the results become

$$\hat\beta = 61.6746, \qquad \hat\alpha = -39.7468.$$

Thus a seemingly small variation in the data has a real effect.

See also

- Deming regression — simple linear regression with errors measured non-vertically
- Linear segmented regression
- Proofs involving ordinary least squares — derivation of all formulas used in this article in general multidimensional case

References

1. ^ Kenney, J. F. and Keeping, E. S. (1962) "Linear Regression and Correlation." Ch. 15 in *Mathematics of Statistics*, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand, pp. 252-285

External links

- Wolfram MathWorld's explanation of Least Squares Fitting, and how to calculate it (http://mathworld.wolfram.com/LeastSquaresFitting.html)

Retrieved from "http://en.wikipedia.org/w/index.php?title=Simple_linear_regression&oldid=601153008"
Categories: Regression analysis | Estimation theory | Parametric statistics

- This page was last modified on 25 March 2014 at 05:44.