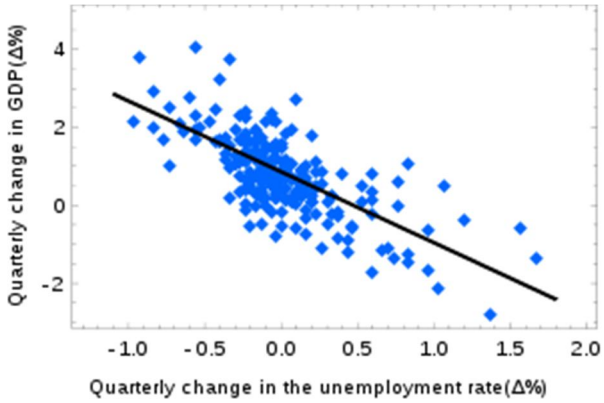
Simple linear regression

From Wikipedia, the free encyclopedia
Jump to: navigation, search

This article **does not cite any references or sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(December 2009)*



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.

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.

The adjective *simple* refers to the fact that this regression is one of the simplest in statistics. The fitted line has the slope 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

[hide]

[edit] Fitting the regression line

Suppose there are n data points $\{y_i, x_i\}$, where $i = 1, 2, \ldots, n$. 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: such a line that minimizes the sum of squared residuals of the linear regression model. In other words, numbers α and β solve the following minimization problem:

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

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 are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} y_j}{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\operatorname{Cov}[x,y]}{\operatorname{Var}[x]} = r_{xy}\frac{s_y}{s_x},$$

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

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. Horizontal bar over a variable means the sample average of that variable. For example: $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$.

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.

[edit] Properties

1. The line goes through the "center of mass" point (x, 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: $\sum_{i=1}^{n} x_i\hat{\varepsilon}_i = 0.$
4. 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.

[edit] 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 $\hat{\beta} = (\overline{xy})/(\overline{x^2})$.

[edit] Linear regression with non-uniform errors

If the y_i have independent, non-uniform variances σ_i^2, then the function Q above becomes

$$Q(\alpha, \beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2/\sigma_i^2$$

However, the OLS estimators for α and β are still given by the same equations as above,

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x},$$

except with the averages computed as weighted means of the appropriate variable combinations.

[edit] Total least squares method

The above equations assume that $\{x_i\}$ data are known exactly whereas $\{y_i\}$ data have random distribution. In case that both $\{x_i\}$ and $\{y_i\}$ are random we can minimize the orthogonal distances from the observations to the regression line (total least squares method). Here we will show equation for simple linear regression model without the intercept term $y = \beta x$ under the assumption that x and y have equal variances. For general case please see Deming regression.

The orthogonal distance from the observation point $\{y_i, x_i\}$ to the regression line $y = \beta x$ is:

$$D_i = \frac{|\beta x_i - y_i|}{\sqrt{\beta^2 + 1}},$$ where $|\beta x_i - y_i|$ denotes absolute value.

We will minimize sum of squared distances:

$$S(\beta) = \sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} \frac{(\beta x_i - y_i)^2}{\beta^2 + 1}$$

This can be solved by searching $\hat{\beta}$ for which derivation $\frac{dS(\beta)}{d\beta}$ is zero.

This yields the equation

$$\hat{\beta}^2 \sum_{i=1}^{n} x_i y_i + \hat{\beta} \sum_{i=1}^{n} (x_i^2 - y_i^2) - \sum_{i=1}^{n} x_i y_i = 0$$

Which has the solution:

$$\hat{\beta}_{1,2} = \frac{-\sum_{i=1}^{n}(x_i^2 - y_i^2) \pm \sqrt{[\sum_{i=1}^{n}(x_i^2 - y_i^2)]^2 + 4[\sum_{i=1}^{n} x_i y_i]^2}}{2\sum_{i=1}^{n} x_i y_i}$$

Please note that there are two different solutions (note $\pm$ in the equation above). One solution represent minimum of $S(\beta)$ function and the other one maximum of $S(\beta)$ function. The corresponding lines are orthogonal to each other. For $x_i > 0$ and $y_i > 0$ for all i use + sign to get the slope of the regression line.

It can be shown that if you swap x_i and y_i you will get slope $\frac{1}{\hat\beta}$. This is not the case for simple linear regression without the intercept term which is using ordinary least squares method.

[edit] 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 do not tell us how precise the estimates are. That is, how much the estimators $\hat\alpha$ and $\hat\beta$ can deviate from the "true" values of α and β. The latter question is answered by the *confidence intervals* for the regression coefficients.

In order to construct the confidence intervals usually one of the two possible assumptions is made: either that the errors in the regression are normally distributed (the so-called *classic regression* assumption), or that the number of observations n is sufficiently large so that the actual distribution of the estimators can be approximated using the Central Limit Theorem.

[edit] 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. 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

$$t = \frac{\hat\beta - \beta}{s_{\hat\beta}} \sim t_{n-2}, \quad \text{where} \quad s_{\hat\beta} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} \hat\varepsilon_i^{\,2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

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

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

$$\beta \in \left[\ \hat\beta - s_{\hat\beta} t^*_{n-2},\ \hat\beta + s_{\hat\beta} t^*_{n-2}\ \right] \text{ at confidence level } (1-\gamma),$$

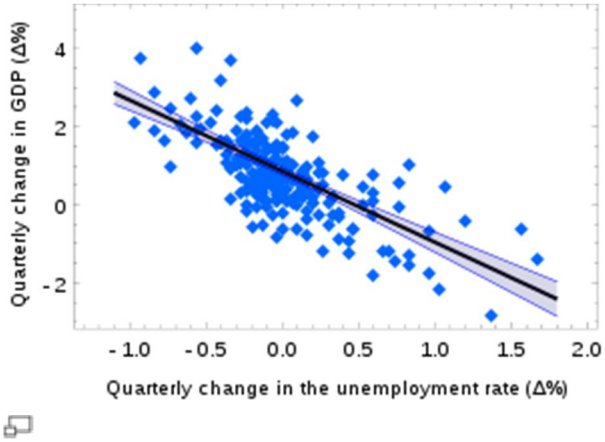
where t^*_{n-2} is the $(1-\gamma/2)$-th quantile of the t_{n-2} distribution. For example, if $\gamma = 0.05$ then the confidence level is 95%.

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

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



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

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 $\hat{\alpha} = 0.859$ and $\hat{\beta} = -1.817$. The 95% confidence intervals for these estimates are

$$\alpha \in [0.76,\ 0.96], \quad \beta \in [-2.06,\ -1.58]$$ with 95% confidence.

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

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

[edit] 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-*t* distribution is replaced with the quantile q^* of the standard normal distribution. Occasionally the fraction $1/(n-2)$ is replaced with $1/n$. When n is large such change does not alter the results considerably.

[edit] Numerical example

As an example we shall consider 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, we will not do so and fit the simple linear regression 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, and we start by calculating the following five sums:

$$S_x = \sum x_i = 24.76, \quad S_y = \sum y_i = 931.17$$
$$S_{xx} = \sum x_i^2 = 41.0532, \quad S_{xy} = \sum x_i y_i = 1548.2453, \quad S_{yy} = \sum y_i^2 = 5849$$

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

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

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

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

[edit] Beware

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 2.54cm, 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

$$\hat{\beta} = 61.6746$$
$$\hat{\alpha} = -39.7468$$

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

[edit] See also

- Proofs involving ordinary least squares — derivation of all formulas used in this article in general multidimensional case;
- Deming regression — orthogonal simple linear regression.
- Linear segmented regression