Linear Regression

03/24/2025

Linear regression is to estimate the coefficients of a linear relation between two variables, x and y. The linear relation is represented by the equation below:

$$y = \beta_0 + \beta_1 x$$

where x is the explanatory variable, y the response variable, β_0 and β_1 the coefficients.

Linear Least Squares

When a group of n measurements of y_i in response to n x_i is collected, the above linear relation no longer holds precisely because of the presence of factors that affect y_i but are not directly in response to x_i. The error term, ϵ_i, also called noise, has a mean of zero and variance σ^2. Therefore, for each y_i, the relation with x_i becomes:

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

The sum of the error term square, $\sum_{i=1}^{n} \epsilon_i^2$, or residual sum of squares (RSS) is:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The estimation of β_0 and β_1 is to minimize the residual error term, RSS, by setting its partial derivatives with respect to β_0 and β_1 to be zero and solve for them.

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

The estimates for the coefficients from the equation above are:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

This method of estimating $\hat{\beta}_0$ and $\hat{\beta}_1$ is called *linear least squares* since the residual sum of squares (RSS) is minimized as the estimation.

Confidence Intervals

The estimate for σ^2 is:

$$\hat{S}_e^2 = \frac{RSS}{n-2}$$

where $n-2$ is the degrees of freedom since two β parameters are estimated using the measurement data. The estimates for the variance of β_1 and β_0 are:

$$\hat{S}_{\beta_1}^2 = \frac{\hat{S}_e^2/n}{\overline{x^2} - \overline{x}^2}$$

$$\hat{S}_{\beta_0}^2 = \hat{S}_{\beta_1}^2 \overline{x^2}$$

For a large n, the central limit theorem allows to construct the confidence intervals for the estimated β.

$$(\widehat{\beta}_i - \beta_i)/S_{\beta_i} \sim t_{n-2}$$

Therefore, the 100(1-α)% confidence intervals can be written as:

$$\hat{\beta}_0 \pm t_{n-2}(\alpha/2) \times S_{\beta_0}$$
$$\hat{\beta}_1 \pm t_{n-2}(\alpha/2) \times S_{\beta_1}$$

Likewise, for hypothesis testing of $H_0 : \beta_i = 0$ i.e., x_i having no predictive power over y_i, the p-value of the test is computed as:

$$p_{\hat{\beta}_i} = P\left(t = \hat{\beta}_i/S_{\beta_i}\right)$$

The prediction interval for $\hat{y}_i$ is as follows:

$$\hat{y}_i \pm t_{n-2}(\alpha/2) S_e \sqrt{1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2/n}{\overline{x^2} - \overline{x}^2}}$$

Correlation Coefficient

The estimated linear regression relation can be written as an expression using the correlation coefficient, r.

$$\frac{\hat{y}-\overline{y}}{\sqrt{S_{yy}}}=r\frac{\hat{x}-\overline{x}}{\sqrt{S_{xx}}}$$

Since $r \leq 1$, this expression can be interpreted as follows. When $\hat{x}$ is one standard deviation from its mean, the response variable $\hat{y}$ is less than or equal to its standard deviation, discounted by r which is always less than or equal to one.

R^2 can be expressed as follows:

$$R^2=\frac{S_y^2-S_\epsilon^2}{S_y^2}$$

It is the ratio of the difference between the variance of the response variable (y) and the variance of the residual error (ϵ) to the variance of the response variable. It therefore is the portion of the variability in the response variable that is in response to the explanatory variable (x), not to the error term (ϵ). As such, R^2 often is called the coefficient of determination and used to measure the strength of the linear relation derived from the least squares method.

Concluding Remarks

Linear regression is often accomplished by the linear least squares method that provides point estimates of the coefficients in the linear relation (β_i). The confidence interval of these estimates can be constructed from the variance estimates and the critical value of t distribution with the degrees of freedom of $n-2$ at a given confidence level α. "p-values" of the estimated coefficients are often provided by a linear regression tool, *e.g.*, a spreadsheet; their computations are merely the probabilities of the t scores with the null hypothesis that these coefficients are zero. Prediction intervals are frequently used in the statistical plots to show the error band of the estimated relation. Again, the band is related to the RSS and the critical value of the t distribution at a given significance level.

The correlation coefficient has a unique meaning in the variation of the response variable from its mean in terms of standard deviation. It is always less than or equal to that of the explanatory variable in its deviation from the mean. The coefficient of determination (R^2) has been used broadly as the output of technical curve fitting. However, the rationale for its use has not been clearly explained in the literature.