Linear Regression

Robert Haas

Documentation • February 15, 2026 • Version 1.0

••• THIS IS STILL A DRAFT •••

PDF generated with LaTeX

Project page: https://github.com/Haasrobertgmxnet/HeatConduction
Contact: Haasrobert@gmx.net

Abstract

This documentation presents ...

Contents

1 Least Squares of correlated Data

1.1 Statement of the Problem and its Solution

Given a data set $\{(x_i, y_i) \in \mathbb{R}^2,\ i \in \{1,\ldots,n\}\}$ we ask for an approximate functional dependency of the form $y_i \approx f_i = f(x_i)$ for all $i \in \{1,\ldots,n\}$. This means, f is a mathematical function assigning each element x of an interval $I \subset \mathbb{R}$ an unique $y \in \mathbb{R}$. For the rest of this documentation we assume that f is a linear function, i.e. there exist numbers β_0 and β_1 in $\mathbb{R}$ such that $f(x) = \beta_0 + \beta_1 x$. To achieve an optimal f with optimal numbers β_0 and β_1 usually a least-square approach is used:

$$\min_{\beta_0,\beta_1} F(\beta_0,\beta_1) \text{ with } F(\beta_0,\beta_1) = \frac{1}{2}\sum_{i=1}^{n} |\beta_0 + \beta_1 x_i - y_i|^2. \tag{1}$$

For optimal $\hat{\beta}_0$ and $\hat{\beta}_1$ the necessary conditions read as

$$\frac{\partial F}{\partial \beta_1} = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) x_i = 0 \text{ and } \frac{\partial F}{\partial \beta_0} = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - y_i) = 0. \tag{2}$$

Now we need the following definitions:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \text{ the arithmetic mean of } \{x_i\}_{i=1}^{n}, \tag{3}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \text{ the arithmetic mean of } \{y_i\}_{i=1}^{n}, \tag{4}$$

$$\sigma_x = \sqrt{\nu_n \sum_{i=1}^{n} (x_i - \bar{x})^2}, \text{ the standard deviation of } \{x_i\}_{i=1}^{n}, \tag{5}$$

$$\sigma_y = \sqrt{\nu_n \sum_{i=1}^{n} (y_i - \bar{y})^2}, \text{ the standard deviation of } \{y_i\}_{i=1}^{n}, \tag{6}$$

$$\varrho_{x,y} = \frac{\nu_n \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y}, \text{ the correlation coefficient of } \{x_i\}_{i=1}^{n} \text{ and } \{y_i\}_{i=1}^{n}. \tag{7}$$

Here ν_n is either $1/(n-1)$ if $\{(x_i, y_i)\}_{i=1}^{n}$ is a sample, and $\nu_n = 1/n$, else. With this definitions the solutions $\hat{\beta}_0$ and $\hat{\beta}_1$ of (2) are

$$\hat{\beta}_1 = \frac{\sigma_y}{\sigma_x}\varrho_{x,y} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{8}$$

$\hat{\beta}_1 = \sigma_y \varrho_{x,y}/\sigma_x$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

1.2 The Coefficient of Determination

A essential question is how good such an OLS model performs. Of course, one wants minimal sum of squares $\sum_{i=1}^{n} (f_i - y_i)^2$. On the other hand large deviations in y_i may prevent a good explanatory power of the model. The answer will be given by the coefficient of determination. To go into more detail we need some more definitions:

$$\begin{aligned} SS_{\text{tot}} &= \sum_{i=1}^{n}(y_i-\bar{y})^2, \text{ the total sum of squares },\\ SS_{\text{reg}} &= \sum_{i=1}^{n}(f_i-\bar{y})^2, \text{ the explained sum of squares },\\ SS_{\text{res}} &= \sum_{i=1}^{n}(f_i-y_i)^2, \text{ the residual sum of squares }. \end{aligned}$$

Theorem 1.1. *The equation $SS_{\text{tot}} = SS_{\text{reg}} + SS_{\text{res}}$ is true.*

Proof. There is

$$\begin{aligned} SS_{\text{tot}} &= \sum_{i=1}^{n}(y_i-\bar{y})^2\\ &= \sum_{i=1}^{n}((y_i-f_i)+(f_i-\bar{y}))^2\\ &= SS_{\text{reg}} + SS_{\text{res}} - 2\sum_{i=1}^{n}(f_i-y_i)(f_i-\bar{y}) \end{aligned}$$

It remains to show $\sum_{i=1}^{n}(f_i-y_i)(f_i-\bar{y})=0$. We have

$$\begin{aligned} \sum_{i=1}^{n}(f_i-y_i)(f_i-\bar{y}) &= \sum_{i=1}^{n}\Big((f_i-y_i)(\hat{\beta}_0+\hat{\beta}_1x_i)-(f_i-y_i)\bar{y}\Big)\\ &= (\hat{\beta}_0-\bar{y})\sum_{i=1}^{n}(f_i-y_i)+\hat{\beta}_1\sum_{i=1}^{n}(f_i-y_i)x_i. \end{aligned}$$

Finally we have

$$\begin{aligned} \sum_{i=1}^{n}(f_i-y_i) &= \frac{\partial F}{\partial \beta_0} = \sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1x_i-y_i)=0,\\ \sum_{i=1}^{n}(f_i-y_i)x_i &= \frac{\partial F}{\partial \beta_1} = \sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1x_i-y_i)x_i=0, \end{aligned}$$

as from (2). That is $\sum_{i=1}^{n}(f_i-y_i)(f_i-\bar{y})$ and $SS_{\text{tot}} = SS_{\text{reg}} + SS_{\text{res}}$. □

The coefficient of determination R^2 is given by

$$R^2 = 1-\frac{SS_{\text{res}}}{SS_{\text{tot}}} = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}.$$

Theorem 1.2. *For the choice (8) in the OLS problem (1) the equation $\varrho^2_{x,y} = R^2$ is true.*

Proof. There is

$$\begin{aligned}\frac{SS_{\text{reg}}}{SS_{\text{tot}}} &= \frac{\sum_{i=1}^{n}(\hat{\beta}_0+\hat{\beta}_1 x_i-\bar{y})^2}{SS_{\text{tot}}}\\ &= \frac{\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\\ &= \frac{\hat{\beta}_1^2\sigma_x^2}{\sigma_y^2}\\ &= \left(\frac{\sigma_y}{\sigma_x}\varrho_{x,y}\right)^2\frac{\sigma_x^2}{\sigma_y^2}=\varrho_{x,y}^2\end{aligned}$$

□

1.3 Confidence Intervals

For the statistical estimations it is convenient to give a confidence interval. So it is reasonable to calculate a confidence interval for the slope $\hat{\beta}_1$. [2, p. 161] and [1, p. 185] give equivalent formulas for such a confidence interval.

References

[1] J. Bortz, *Statistik für Sozialwissenschaftler* 5th ed., Springer-Verlag Berlin Heidelberg, 1999.

[2] M. Sachs, *Wahrscheinlichkeitsrechnung und Statistik* 6th ed., Verlag Carl Hanser, München, 2021.

[3] Code samples used in this work: `https://github.com/Haasrobertgmxnet/LinearRegression`