Simple Linear Regression: Three Sets of Basic Equations

Glen Eric Meyer

January 25, 2023

Contents

1 Introduction **1**

2 The Equations **1**

2.1 Least (Vertical) Squares 1

2.2 A Simpler Alternative 2

2.3 The Relation Between Lease Squares Regression and Correlation Analysis 2

1 Introduction

Simple linear regression concerns the following simple statistical model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, ..., n,$$

where the y_i are quantities we want to predict or understand (the response or dependent variables), the e_i are independent random variables, and the x_i, the predictor or independent variables, are assumed to be fixed quantities (that is, *not* random); in addition, $\forall i, 1 \leq i \leq n$,

$$E[e_i] = 0,$$

and

$$Var[e_i] = \sigma^2.$$

Note that to perform a least squares analysis we do not, strictly speaking, need a *statistical* model, but if we want to assess how well our model fits the data, especially in the presence of noise, we must at least specify the mean and variance of the error term and, in many cases, also specify a full probability distribution (usually a Gaussian distribution) for the error term.

2 The Equations

2.1 Least (Vertical) Squares

Here are equations for estimates of the regression coefficients, the intercept β_0 and the slope β_1 in terms of the data (x_i, y_i):

$$\hat{\beta}_0 = \frac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

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

2.2 A Simpler Alternative

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

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

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

2.3 The Relation Between Lease Squares Regression and Correlation Analysis

Least squares regression is closely related to the analysis of the correlation between the y_i's and the x_i's and can be formulated in terms of such an analysis. To see how, first define some new variables:

$$S_{xx} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

With these definitions, the correlation coefficient between the y_i's and the x_i's is

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

In terms of the above, here are the equations for estimates of the regression coefficients:

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

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Note that the equation for $\hat{\beta}_1$ together with the equation for the correlation coefficient r give us:

$$\hat{\beta}_1 = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$

and

$$r = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$