Simple Linear Regression

Contents

Introduction

Goal

The goal of simple linear regressions is to model the relationship between two variables x and y by fitting a linear (affine more generally) function f of the form:

$$y = f(x) = \beta_0 + \beta_1 x \tag{1}$$

where $x \in \mathbb{R}$ is the independent variable, $y \in \mathbb{R}$ is the dependent variable, $\beta_0 \in \mathbb{R}$ is the intercept, and $\beta_1 \in \mathbb{R}$ is the coefficient of x.

Underlying Relationship Formulation

There could be underlying unobserved deviations from the equation (1) which are called errors. Thus, for any data pair (x_i, y_i), the underlying true relationship between x_i and y_i can be described by involving the error term ϵ_i into the equation:

$$y_i = f(x_i) = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2}$$

This relationship between the true (but unobserved) underlying parameters β_0 and β_1 and the data pairs is called a linear regression model.

Prediction

It is indispensable to find sufficiently good estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1 in simple linear regression problems, so that the predicted value of y is given by:

$$\hat{y} = \hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3}$$

for any given value of x.

Estimations of Parameters

Goal

In order to find a good estimation of y with any given value of x, finding decent estimates of β_0 and β_1 is essential. I.e., find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that minimizes the residuals $\{\hat{\epsilon}_i\}_{i=1}^n$ which are the differences between the actual value of y_i and the predicted value of y_i (i.e. $\hat{y}_i$):

$$\begin{aligned}
\hat{\epsilon}_i &= y_i - \hat{y}_i && (4)\\
&= \beta_0 + \beta_1 x_i + \epsilon_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \text{ by (2) and (3)} && (5)\\
&= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) x_i + \epsilon_i && (6)
\end{aligned}$$

Note the difference between a residual $\hat{\epsilon}_i$ and an error ϵ_i, see Errors and Residuals.

Ordinary Least Square Estimation

Introduction

The most commonly-used way to estimate the parameters is by least-square regression/estimation or ordinary least square (OLS) estimation. OLS chooses the parameters of a linear function of a set of explanatory variables by the principle of least squares. By Gauss-Markov Theorem, under Gauss-Markov assumptions, OLS estimation yields the best linear unbiased estimator (BLUE), i.e. it is the most efficient estimator (i.e. has lowest variance) among all linear unbiased estimators.

Estimators

The OLS estimation yields $\hat{\beta}_{0OLS}$ and $\hat{\beta}_{1OLS}$ by minimizing the sum of Euclidean distance of y_i and $\hat{y}_i$ which is just the sum of squared residuals $\hat{\epsilon}_i$ as described in equation (4).

The sum of squared residuals with respect of the parameters β_0 and β_1 can be denoted as:

$$E_{OLS}(\beta_0, \beta_1) = \sum_{i=1}^{n} \hat{\epsilon_i}^2 \tag{7}$$
$$= \sum_{i=1}^{n} (y_i - \hat{y_i})^2 \tag{8}$$
$$= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{9}$$

Note that the β_0 and β_1 here are not the the real value of the parameters but variables to be optimized over.

Thus, the OLS estimator of β_0 is given by:

$$\hat{\beta}_{0OLS} = \underset{\beta_0}{\operatorname{argmin}} E_{OLS}(\beta_0, \beta_1) \tag{10}$$

The analytical solution of $\hat{\beta}_{0OLS}$ can be found by applying the derivative test.

First-derivative test:

$$\frac{\partial E_{OLS}(\beta_0, \beta_1)}{\partial \beta_0} = \frac{\partial \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{\partial \beta_0}, \text{ by (9)} \tag{11}$$
$$= -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \tag{12}$$
$$= -2 (\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i) \tag{13}$$

Set $\frac{\partial E_{OLS}}{\partial \beta_0} = 0$ to get extrema of E_{OLS} with respect to β_0, denoting the value of β_0 at the extrema by $\beta_{0extrema}$:

$$\implies \quad -2 (\sum_{i=1}^{n} y_i - n\beta_{0extrema} - \beta_1 \sum_{i=1}^{n} x_i) = 0 \tag{14}$$
$$\implies \quad \beta_{0extrema} = -\frac{\beta_1 \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i}{n} \tag{15}$$
$$\implies \quad \beta_{0extrema} = \bar{y} - \beta_1 \bar{x} \tag{16}$$

where $\bar{x} = \sum_{i=1}^{n} x_i$ is the average of x_i's and $\bar{y} = \sum_{i=1}^{n} y_i$ is the average of y_i's.

Note that for $\hat{\beta}_{0OLS} = \beta_{0extrema}$, it is necessary to check $\frac{\partial^2 E_{OLS}}{\partial \beta_{0extrema}^2} > 0$ (i.e. second-derivative test). However, from equation (9), it is obvious that E_{OLS} is a quadratic function opens upwards with respect of $\hat{\beta}_0$, which ensures that $\hat{\beta}_{0OLS} = \beta_{0extrema} = \operatorname{argmin}_{\beta_0} E_{OLS}(\beta_0, \beta_1)$. Therefore,

$$\hat{\beta}_{0OLS} = \bar{y} - \beta_1 \bar{x} \tag{17}$$

The OLS estimator of β_1 is given by:

$$\hat{\beta}_{1OLS} = \underset{\beta_1}{\operatorname{argmin}} E_{OLS}(\hat{\beta}_{0OLS}, \beta_1) \tag{18}$$

s

Again, the β_1 here is not the the real value of the parameter but a variable to be optimized over.

The analytical solution of $\hat{\beta}_{1OLS}$ can be found by applying the derivative test.

First-derivative test:

$$\begin{aligned}
\frac{\partial E_{OLS}(\hat{\beta}_{0OLS}, \beta_1)}{\partial \beta_1} &= \frac{\partial \sum_{i=1}^{n} (y_i - \hat{\beta}_{0OLS} - \beta_1 x_i)^2}{\partial \beta_1}, \text{ by (9)} &(19)\\
&= \frac{\partial \sum_{i=1}^{n} (y_i - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i)^2}{\partial \beta_1} &(20)\\
&= \frac{\partial \sum_{i=1}^{n} (y_i - \bar{y} + \beta_1(\bar{x} - x_i))^2}{\partial \beta_1} &(21)\\
&= 2\sum_{i=1}^{n} (\bar{x} - x_i)(y_i - \bar{y} + \beta_1(\bar{x} - x_i)) &(22)
\end{aligned}$$

Set $\frac{\partial E_{OLS}}{\partial \beta_1} = 0$ to get extrema of E_{OLS} with respect to β_1, denoting the value of β_1 at the extrema by $\beta_{1extrema}$:

$$\begin{aligned}
&\implies 2\sum_{i=1}^{n} [(y_i - \bar{y})(\bar{x} - x_i) + \beta_{1extrema}(\bar{x} - x_i)^2] = 0 &(23)\\
&\implies \sum_{i=1}^{n} (y_i - \bar{y})(\bar{x} - x_i) + \beta_{1extrema} \sum_{i=1}^{n} (\bar{x} - x_i)^2 = 0 &(24)\\
&\implies \beta_{1extrema} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (\bar{x} - x_i)^2} &(25)
\end{aligned}$$

For the same reason as described above, $\beta_{1extrema} = \operatorname{argmin}_{\beta_1} E_{OLS}(\hat{\beta}_{0OLS}, \beta_1)$. Thus,

$$\hat{\beta}_{1OLS} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (\bar{x} - x_i)^2} \tag{26}$$