

Multiple Linear Regression

Adam Lee

I. INTRODUCTION

Our definition of linearity yields a significant amount of flexibility for our regression models. We can extend our definition for simple linear regression to form a regression model with multiple regressors each of which does not itself need to be linear but instead is accompanied by a linear coefficient. That is we can fit a model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon, \quad (1)$$

where y is our response variable and x_i for $i = 1, \dots, n$ are our regressors. It must be noted the term linear is used here since eq. (1) is a linear function of the unknown parameters β_i . This means we can choose our regressor terms freely. More complex models may allow for a polynomial fit.

Example 1: Consider the polynomial model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \quad (2)$$

If we let $x_i := x^i$ then this model takes the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \quad (3)$$

and is a linear regression model with three unknown regression coefficients and two regressors. It is still linear in our coefficients β_i however allows a much more flexible model.

We can extend our definition of a multiple linear regression model further to include interaction terms. For example, consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2. \quad (4)$$

If we rewrite this model such that $\beta_{12} x_1 x_2 = \beta_3 x_3$ then we have a model of a familiar form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

II. MULTIPLE LINEAR REGRESSION

Let us focus on the generic case with model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon. \quad (5)$$

This model describes an $n + 1$ dimension hyperplane, as opposed to the regression line we constructed for the simple case. The parameters β_i are our regression coefficients and in practice these, as well as the variance of the error ε , are unknown. The particular parameter β_j represents the expected change in y per unit change in x_j under the assumption that all other regressors x_k , ($k \neq j$) are constant. Figure 1 provides graphic representation of a two-regressor, non-interacting model.

Example 2: Consider a model where our expected error is zero

$$E(y) = 10 + 15x_1 + 4x_2. \quad (6)$$

We can examine the hyperplane this model describes.

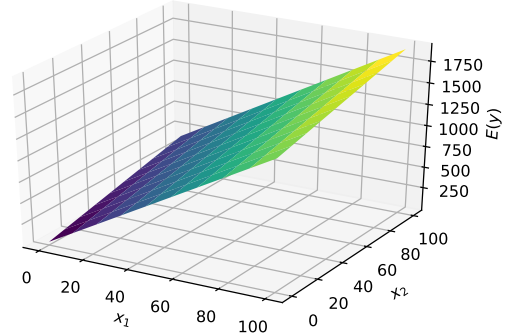


Fig. 1: Hyperplane plot for eq. (6)

Let us assume we have a data-set $\mathcal{X} = \{y_i, x_{i1}, \dots, x_{in}\}_{i=0}^k$ where $n > k$. Here, y_i denotes the i -th observed response and x_{ij} denotes the observed value x_i for regressor x_j . Then, we have a system of model equations

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in} \quad i = 0, \dots, k. \quad (7)$$

We can then assume matrix notation for the entire set such that

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \quad (8)$$

where

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{pmatrix}, \quad (9)$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{01} & \cdots & x_{0n} \\ 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \cdots & x_{kn} \end{pmatrix}, \quad (10)$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \quad (11)$$

and

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{pmatrix}. \quad (12)$$

The task then becomes fitting this linear model such that the coefficients $\boldsymbol{\beta}$ minimize the error term $\boldsymbol{\varepsilon} = \mathbf{y} - X\boldsymbol{\beta}$. A common method for determining these parameters is the method of least-squares, which we have already come across for the simple linear regression case. The same principles can be extended to multiple linear regression. Let us set $p = k + 1, q = n + 1$ moving forward for convenience.

III. METHOD OF LEAST-SQUARES

Let us define the multivariate least-squares criterion function

$$\begin{aligned} \mathbb{E}(\boldsymbol{\beta}) &:= \sum_{i=0}^k \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta}) \\ &= \mathbf{y}^T \mathbf{y} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta}. \end{aligned} \quad (13)$$

Then, we look for $k + 1$ vector $\hat{\boldsymbol{\beta}}$ such that

$$\arg \min \mathbb{E}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}. \quad (14)$$

To do so we must have

$$\left. \frac{\partial \mathbb{E}}{\partial \boldsymbol{\beta}} \right|_{\hat{\boldsymbol{\beta}}} = -2X^T \mathbf{y} + 2X^T X \hat{\boldsymbol{\beta}} = 0, \quad (15)$$

which we can simplify to

$$X^T X \hat{\boldsymbol{\beta}} = X^T \mathbf{y}. \quad (16)$$

Then, provided the inverse matrix $(X^T X)^{-1}$ exists, our solution is

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}. \quad (17)$$

Let us examine the matrix $X^T X$ in detail. By theorem 1, $X^T X$ is a positive-definite matrix of the form

$$\begin{bmatrix} q & \sum_{i=0}^k x_{i1} & \sum_{i=0}^k x_{i2} & \cdots & \sum_{i=0}^k x_{in} \\ \sum_{i=0}^k x_{i1} & \sum_{i=0}^k x_{i1}^2 & \sum_{i=0}^k x_{i1}x_{i2} & \cdots & \sum_{i=0}^k x_{i1}x_{in} \\ \sum_{i=0}^k x_{i2} & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \sum_{i=0}^k x_{in} & \sum_{i=0}^k x_{i1}x_{in} & \cdots & \cdots & \sum_{i=0}^k x_{in}^2 \end{bmatrix}$$

APPENDIX

Theorem 1. For any real, invertible matrix $A \in \mathbb{R}^{n \times n}$, the product matrix $A^T A$ is positive definite i.e

$$\mathbf{z}^T A^T A \mathbf{z} > 0 \quad \forall \mathbf{z} \in \mathbb{R}^n \quad (18)$$

Proof.

$$\begin{aligned} \mathbf{z}^T A^T A \mathbf{z} &= (A\mathbf{z})^T (A\mathbf{z}) \\ &= \|A\mathbf{z}\|^2 > 0. \end{aligned} \quad (19)$$

□