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

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon, \tag{1}$$

where y is our response variable and x_i for $i = 1, \ldots, n$ 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

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon. \tag{2}$$

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \tag{3}$$

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

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

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

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon. \tag{5}$$

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

$$E(y) = 10 + 15x_1 + 4x_2. \tag{6}$$

We can examine the hyperplane this model describes.

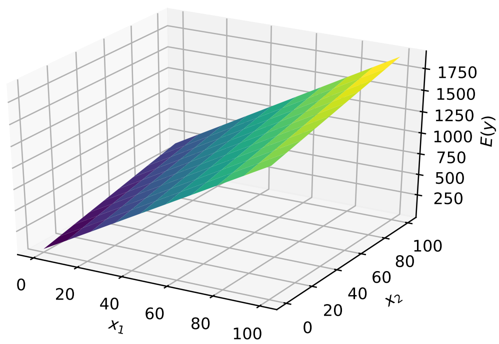

Fig. 1: Hyperplane plot for eq. (6)

Let us assume we have a data-set $\varkappa = \{y_i, x_{i1}, \dots, x_{in}\}_{i=0}^{k}$ 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

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

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

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{8}$$

where

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

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

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

and

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

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

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

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

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

To do so we must have

$$\frac{\partial \mathbb{E}}{\partial \boldsymbol{\beta}}\Big|_{\hat{\boldsymbol{\beta}}} = -2X^{\mathrm{T}}\mathbf{y} + 2X^{\mathrm{T}}X\hat{\boldsymbol{\beta}} = 0, \tag{15}$$

which we can simplify to

$$X^{\mathrm{T}}X\hat{\boldsymbol{\beta}} = X^{\mathrm{T}}\mathbf{y}. \tag{16}$$

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

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

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

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

Inversion of this matrix can be tricky using orthodox methods in practical examples with large data-sets. However, we know enough about this matrix to manipulate it in such a way where this problem is simplified [2]. We can then test our parameters on some set of test points, say $\mathbf{x}^*$ then

$$\hat{\mathbf{y}} = \hat{X}\hat{\boldsymbol{\beta}} = \hat{X}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{y} = H\mathbf{y}, \tag{18}$$

where $\hat{X}$ is the corresponding X for the test points $\mathbf{x}^*$. We call the matrix H the hat matrix of our model, since it puts a hat on $\mathbf{y}$.

IV. Basis Functions

Our generic example considers the construction of an $n+1$ dimension hyperplane, which we can visualise in up to three dimensions as in fig. 1. Let us instead consider examples more akin to example 1 where our model takes the form

$$y = \sum_{i=0}^{n} \beta_n \phi_n(x) + \epsilon, \tag{19}$$

where β_i are our unknown parameters, and ϕ_i are basis functions. Then we can assume matrix notation for our data set $\varkappa$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{20}$$

where

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_k \end{pmatrix}, \tag{21}$$

$$X = \begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \cdots & \phi_n(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_n(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_k) & \phi_1(x_k) & \cdots & \phi_n(x_k) \end{pmatrix}, \tag{22}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}, \tag{23}$$

and

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_k \end{pmatrix}. \tag{24}$$

Then our solution for our parameters is nothing else than

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathbf{y}. \tag{25}$$

A. Basis Function Examples

Our basis functions can assume any form such that each function ϕ_i is real-valued and scalar. Two more popular basis functions are polynomial and Gaussian radial basis functions.

Example 3 (Polynomial Regression)**:** Polynomial regression uses basis functions

$$\boldsymbol{\phi} = \begin{bmatrix} 1 & x & x^2 & \cdots & x^n \end{bmatrix}, \tag{26}$$

then our matrix X would become

$$X = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_k & \cdots & x_k^n \end{pmatrix}. \tag{27}$$

Restricting our basis functions to $\boldsymbol{\phi} = \begin{bmatrix} 1 & x \end{bmatrix}$ gives us simple linear regression.

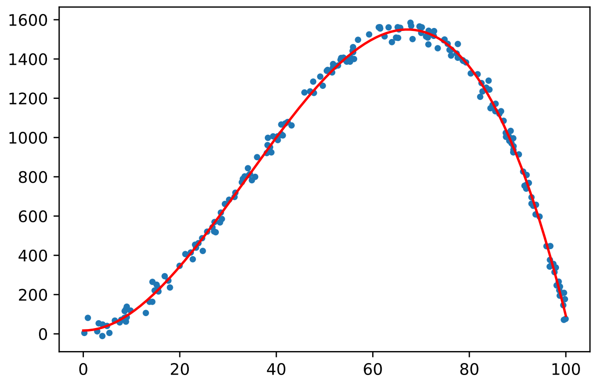

Fig. 2: Example of polynomial regression

Example 4 (Gaussian Radial Basis Functions)**:** Let us now consider Gaussian radial basis functions

$$\phi_i(x) = e^{-\lambda^2(x-\mu_i)^2} \tag{28}$$

where λ is some coefficient representing the variance of our basis functions and $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 & \cdots & \mu_n \end{bmatrix}$ are centres for each Gaussian. Then our predictive function becomes a linear combination of weighted Gaussians where our parameters λ and μ_i determine the smoothness of this function on our test interval. There arises the problem of under or over-fitting our data should we include poor hyper-parameters or attempt to fit too many/few Gaussians to our data.

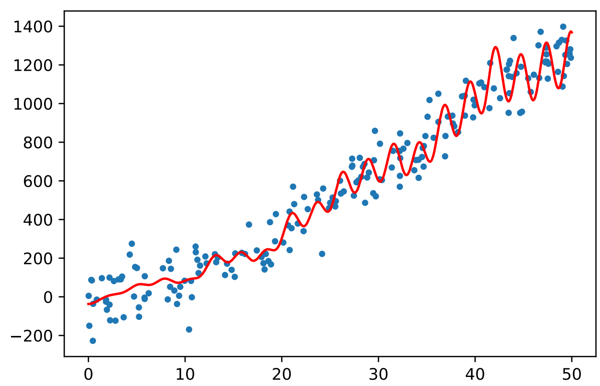

Fig. 3: Example of linear regression using Gaussian radial basis functions

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*

$$\mathbf{z}^T A^T A\mathbf{z} > 0 \ \forall \ \mathbf{z} \in \mathbb{R}^n \tag{29}$$

Proof.

$$\begin{aligned} \mathbf{z}^{\mathrm{T}} A^{\mathrm{T}} A\mathbf{z} &= (A\mathbf{z})^{\mathrm{T}}(A\mathbf{z}) \\ &= ||Az||^2 > 0. \end{aligned} \tag{30}$$

□

Theorem 2. *For a positive-definite matrix A, there exists a lower-triangular matrix U such that*

$$A^{-1} = UU^T \tag{31}$$

Proof. For positive-definite matrix A, we have a Cholesky decomposition such that

$$A = LL^{\mathrm{T}}, \tag{32}$$

then we have

$$AA^{-1} = LL^{\mathrm{T}}A^{-1} = I, \tag{33}$$

that is

$$A^{-1} = (L^{\mathrm{T}})^{-1}L^{-1}. \tag{34}$$

Since L is lower triangular, if we write $(L^{\mathrm{T}})^{-1} = U$ then we have

$$A^{-1} = UU^{\mathrm{T}}, \tag{35}$$

which is a positive-definite symmetric matrix such that we need only compute the upper-triangular elements of UU^{T} in order to know all elements of A^{-1}. □