

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 coefficient which is linearly related to y . That is we can fit a model

$$y = \beta_0 + \beta_1\phi_1(x) + \cdots + \beta_k\phi_k(x) + \varepsilon, \quad (1)$$

where y is our response variable and ϕ_i for $i = 1, \dots, k$ are our basis functions evaluated at our regressor x . 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 such as in example 1.

Example 1: Consider the polynomial model

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \varepsilon. \quad (2)$$

If we let $\phi_i(x) := x^i$ then this model takes the form

$$y = \beta_0 + \beta_1\phi_1(x) + \beta_2\phi_2(x) + \varepsilon, \quad (3)$$

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 for a much more flexible model. The inclusion of more regressors than our simple model presents opportunities for linear regression in higher dimensions [1], while the consideration of basis functions ϕ allows us to consider relationships between the response variable y and the regressors which are not themselves linear.

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

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_{12}x_1x_2, \quad (4)$$

which is still linear in the parameters β_i .

II. MULTIPLE LINEAR REGRESSION

Let us focus on the generic case with model

$$y = \beta_0 + \beta_1x_1 + \cdots + \beta_kx_k + \varepsilon. \quad (5)$$

Here our regressor $\mathbf{x} = [x_1, x_2, \dots, x_k]^T$ is a vector and our basis functions $\phi_i(\mathbf{x}) = x_i$, $i = 1, \dots, k$ simply return each component of the regressor. This model describes a k 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_i , ($i \neq j$) 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 of dimension 2¹ this model describes [1].

¹A plane of dimension d embedded into a space of dimension $d + 1$ is a hyperplane of dimension d .

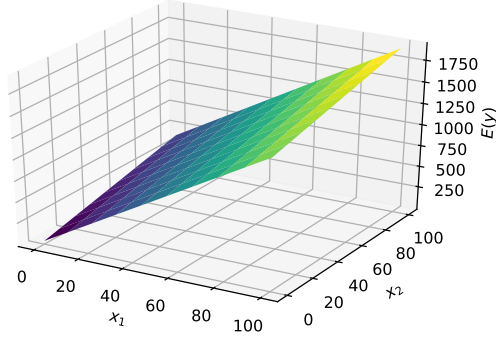


Fig. 1: Hyperplane plot for eq. (6)

Let us assume we have a data-set $\mathcal{X} = \{y_i, x_{i1}, \dots, x_{ik}\}_{i=0}^n$ where $k > n$. 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

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} \quad i = 0, \dots, n. \quad (7)$$

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

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

where

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

$$X = \begin{pmatrix} 1 & x_{01} & \dots & x_{0k} \\ 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad (10)$$

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

and

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

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.

III. METHOD OF LEAST-SQUARES

Let us define the multivariate least-squares criterion function

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

Then, we look for $\hat{\beta} \in \mathbb{R}^{k+1}$ such that

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

To do so we must have

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

which we can simplify to

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

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

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

Let us examine the matrix $X^T X$ in detail. By theorem 1, $X^T X$ is a positive-definite matrix of the form, setting $p = n + 1, q = k + 1$

$$\begin{bmatrix} p^2 & \sum_{i=0}^n x_{i1} & \sum_{i=0}^n x_{i2} & \cdots & \sum_{i=0}^n x_{ik} \\ \sum_{i=0}^n x_{i1} & \sum_{i=0}^n x_{i1}^2 & \sum_{i=0}^n x_{i1}x_{i2} & \cdots & \sum_{i=0}^n x_{i1}x_{ik} \\ \sum_{i=0}^n x_{i2} & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \sum_{i=0}^n x_{ik} & \sum_{i=0}^n x_{i1}x_{ik} & \cdots & \cdots & \sum_{i=0}^n x_{ik}^2 \end{bmatrix}$$

The computational complexity of matrix inversion is of order $\mathbf{O}(n^3)$ for an $n \times n$ matrix. However, the matrix $X^T X$ is positive-definite [2] and therefore we can compute its inverse using Cholesky arithmetic. This process has complexity of order $\mathbf{O}(\frac{1}{2}n^3)$. We can then utilise our parameters on some set of test points, say \mathbf{x}^* then

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

where \hat{X} is the corresponding X for the test points \mathbf{x}^* . We call the matrix H the hat matrix of our model.

IV. BASIS FUNCTIONS

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

$$y = \sum_{i=0}^k \beta_i \phi_i(x) + \epsilon, \quad (19)$$

where β_i are our unknown parameters, and ϕ_i are basis functions. Then we can assume matrix notation for our data set \mathcal{X}

$$\mathbf{y} = X\beta + \epsilon. \quad (20)$$

where

$$\mathbf{y} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad (21)$$

$$X = \begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \cdots & \phi_k(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_k(x_1) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_n) & \phi_1(x_n) & \cdots & \phi_k(x_n) \end{pmatrix}, \quad (22)$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad (23)$$

and

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \quad (24)$$

Then our solution for our parameters is nothing else than

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

A. Basis Function Examples

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

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

$$\boldsymbol{\phi} = [1 \quad x \quad x^2 \quad \cdots \quad x^k]^T, \quad (26)$$

then our matrix X would become

$$X = \begin{pmatrix} 1 & x_0 & \cdots & x_0^k \\ 1 & x_1 & \cdots & x_1^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^k \end{pmatrix}. \quad (27)$$

Restricting our basis functions to $\boldsymbol{\phi} = [1 \quad x]^T$ gives us simple linear regression. In fig. 2, we generated 100 data-points $y_i, i = 1, \dots, 100$ of the form

$$y_i = f(x_i) + \varepsilon, \quad f(x) = 2 + x + x^2 - \frac{1}{100}x^3, \quad \varepsilon \sim \mathcal{N}(0, 1000), \quad (28)$$

which are plotted in blue. The red curve is our regression line computed using MLE fitting a model of the form

$$y = \beta_0 + \beta_1 \phi_1(x) + \beta_2 \phi_2(x) + \beta_3 \phi_3(x), \quad \phi_i(x) = x^i, \quad (29)$$

to our observations.

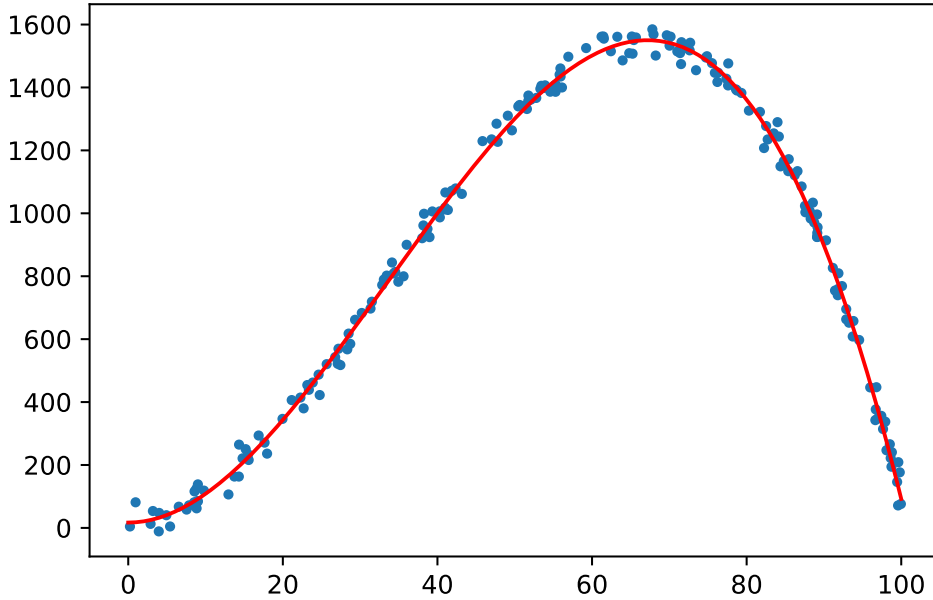


Fig. 2: Example of polynomial regression

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

$$\phi_i(x) = e^{-\lambda^2(x-\mu_i)^2} \quad (30)$$

where λ is some coefficient representing the variance of our basis functions and $\boldsymbol{\mu} = [\mu_1 \ \cdots \ \mu_n]^T$ 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.

Consider fig. 3, here our data was generated in a similar fashion to fig. 2. Our generating function in this case was

$$y_i = f(x_i) + \varepsilon, \quad f(x) = x \sin(0.5x), \quad \varepsilon \sim \mathcal{N}(0, 1). \quad (31)$$

Our basis functions were of the form

$$\phi_i = e^{-0.5(x-i)^2}, \quad i = -10, \dots, 10, \quad (32)$$

that is, we fit a RBF centred at each integer within the domain of our data.

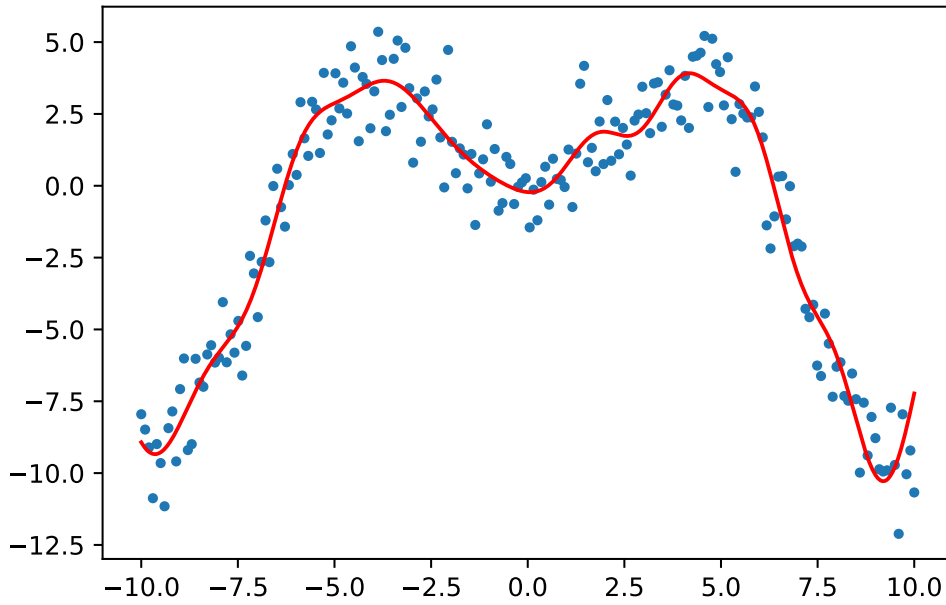


Fig. 3: Example of linear regression using Gaussian radial basis functions

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*

$$\mathbf{z}^T A^T A \mathbf{z} > 0 \quad \forall \mathbf{z} \in \mathbb{R}^n \quad (33)$$

Proof.

$$\begin{aligned} \mathbf{z}^T A^T A \mathbf{z} &= (A\mathbf{z})^T (A\mathbf{z}) \\ &= \|A\mathbf{z}\|^2 > 0. \end{aligned} \quad (34)$$

□

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

$$A^{-1} = U U^T \quad (35)$$

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

$$A = L L^T, \quad (36)$$

then we have

$$A A^{-1} = L L^T A^{-1} = I, \quad (37)$$

that is

$$A^{-1} = (L^T)^{-1} L^{-1}. \quad (38)$$

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

$$A^{-1} = U U^T, \quad (39)$$

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