# The Linear Model and the Method of Least Squares

Given an input vector $\mathbf{x} = (x_1, \ldots, x_p)^T$ we predict the ouput $y(\mathbf{x})$ via the model

$$\widehat{y} = \widehat{\beta}_0 + \sum_{j=1}^{p} x_j \widehat{\beta}_j. \tag{1}$$

where $\{\widehat{\beta}_j\}$ is a set of coefficients. If the constant variable $x_0 = 1$ is included in $\mathbf{x}$ and the coefficients are collected in a column vector $\widehat{\boldsymbol{\beta}}$, we can write the linear model compactly as an inner product,

$$\widehat{y} = \mathbf{x}^T \widehat{\boldsymbol{\beta}}.$$

If we have a set of training data $(\mathbf{x}_i, \; y_i)$ for $i = 1, \ldots, N$, we can pick the coefficients $\widehat{\boldsymbol{\beta}}$ that minimizes the residual sum of squares,

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{N} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)^2,$$

$$\widehat{\boldsymbol{\beta}} = \text{argmin } \mathrm{RSS}(\boldsymbol{\beta}).$$

In the following, $X$ will denote a matrix with $N$ rows and $p+1$ columns with row $i$ equal to $\mathbf{x}_i^T$ (note that the entire first column of $X$ consists of 1's). Then the residual sum of squares can be written as

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left( y_i - \sum_{j=0}^{p} X_{ij} \beta_j \right)^2.$$

It's clear that $\mathrm{RSS}(\boldsymbol{\beta})$ is bounded from below (best case scenario is $\mathrm{RSS}(\boldsymbol{\beta}) = 0$) and unbounded from above (the linear model can be made arbitrarily bad, $\mathrm{RSS}(\boldsymbol{\beta}) \to \infty$). Thus, a (possibly not unique) minimum exists, which is found by setting the gradient with respect to $\boldsymbol{\beta}$ to zero,

$$\left[\frac{\partial}{\partial \boldsymbol{\beta}}\mathrm{RSS}(\boldsymbol{\beta})\right]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = 0.$$

The components of the gradient are given by

$$\frac{\partial}{\partial \beta_k}\mathrm{RSS}(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_k}\sum_{i=1}^{N}\left(y_i - \sum_{j=0}^{p} X_{ij}\beta_j\right)^2$$

$$= \sum_{i=1}^{N}\frac{\partial}{\partial \beta_k}\left(y_i - \sum_{j=0}^{p} X_{ij}\beta_j\right)^2 = \sum_{i=1}^{N} 2\left(y_i - \sum_{j=0}^{p} X_{ij}\beta_j\right)\frac{\partial}{\partial \beta_k}\left(y_i - \sum_{j=0}^{p} X_{ij}\beta_j\right)$$

$$= \sum_{i=1}^{N} 2\left(y_i - \sum_{j=0}^{p} X_{ij}\beta_j\right)\left(-\sum_{j=0}^{p} X_{ij}\frac{\partial \beta_j}{\partial \beta_k}\right) = \sum_{i=1}^{N} 2\left(y_i - \sum_{j=0}^{p} X_{ij}\beta_j\right)\left(-\sum_{j=0}^{p} X_{ij}\delta_{jk}\right)$$

$$= \sum_{i=1}^{N} 2\left(y_i - \sum_{j=0}^{p} X_{ij}\beta_j\right)(-X_{ik}) = -2\sum_{i=1}^{N}\left(X_{ik}y_i - X_{ik}\sum_{j=0}^{p} X_{ij}\beta_j\right)$$

$$= -2\left(\sum_{i=1}^{N} X_{ki}^T y_i - \sum_{i=1}^{N} X_{ki}^T\sum_{j=0}^{p} X_{ij}\beta_j\right)$$

The vector form of this expression is

$$\frac{\partial}{\partial \boldsymbol{\beta}}\mathrm{RSS}(\boldsymbol{\beta}) = -2\big(X^T\mathbf{y} - X^T X\boldsymbol{\beta}\big).$$

Thus, the vector $\widehat{\boldsymbol{\beta}}$ satisfies

$$X^T\mathbf{y} = X^T X\widehat{\boldsymbol{\beta}}.$$

If the matrix $X^T X$ is invertible, the unique minimum is given by

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T \mathbf{y},$$

which is the set of coefficients that we plug into the linear model in Eq. (1).