Matrix Notation for Multiple Linear Regression

This document provides the details for the matrix notation for multiple linear regression. We assume the reader has familiarity with some linear algebra. Please see Chapter 1 of *An Introduction to Statistical Learning* for a brief review of linear algebra.

Introduction

Suppose we have n observations. Let the i^{th} be $(x_{i1}, \ldots, x_{ip}, y_i)$, such that $x_{i1}, \ldots, x_{ip}$ are the explanatory variables (predictors) and y_i is the response variable. We assume the data can be modeled using the least-squares regression model, such that the mean response for a given combination of explanatory variables follows the form in (1).

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

We can write the response for the i^{th} observation as shown in (2)

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{2}$$

such that ϵ_i is the amount y_i deviates from $\mu\{y|x_{i1}, \ldots, x_{ip}\}$, the mean response for a given combination of explanatory variables. We assume each $\epsilon_i \sim N(0, \sigma^2)$, where σ^2 is a constant variance for the distribution of the response y for any combination of explanatory variables $x_1, \ldots, x_p$.

Matrix Representation for the Regression Model

We can represent the (1) and (2) using matrix notation. Let

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{3}$$

Thus,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

Therefore the estimated response for a given combination of explanatory variables and the associated residuals can be written as

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{4}$$

Estimating the Coefficients

The least-squares model is the one that minimizes the sum of the squared residuals. Therefore, we want to find the coefficients, $\hat{\boldsymbol{\beta}}$ that minimizes

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{5}$$

where $\mathbf{e}^T$, the transpose of the matrix $\mathbf{e}$.

$$(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\hat{\boldsymbol{\beta}} - (\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}) \tag{6}$$

Note that $(\mathbf{Y^T X}\hat{\boldsymbol{\beta}})^T = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y}$. Since these are both constants (i.e. 1×1 vectors), $\mathbf{Y^T X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y}$. Thus, (6) becomes

$$\mathbf{Y}^T\mathbf{Y} - 2\mathbf{X}^T\hat{\boldsymbol{\beta}}^T\mathbf{Y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \tag{7}$$

Since we want to find the $\hat{\boldsymbol{\beta}}$ that minimizes (5), will find the value of $\hat{\boldsymbol{\beta}}$ such that the derivative with respect to $\hat{\boldsymbol{\beta}}$ is equal to 0.

$$\begin{aligned}
\frac{\partial \mathbf{e}^T\mathbf{e}}{\partial \hat{\boldsymbol{\beta}}} &= \frac{\partial}{\partial \hat{\boldsymbol{\beta}}}(\mathbf{Y}^T\mathbf{Y} - 2\mathbf{X}^T\hat{\boldsymbol{\beta}}^T\mathbf{Y} + \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}) = 0 \\
&\Rightarrow -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \\
&\Rightarrow 2\mathbf{X}^T\mathbf{Y} = 2\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&\Rightarrow \mathbf{X}^T\mathbf{Y} = \mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&\Rightarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} \\
&\Rightarrow (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{I}\hat{\boldsymbol{\beta}}
\end{aligned} \tag{8}$$

Thus, the estimate of the model coefficients is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

Variance-covariance matrix of the coefficients

We will use two properties to derive the form of the variance-covariance matrix of the coefficients:

1. $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I$
2. $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\epsilon$

First, we will show that $E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \sigma^2 I$

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = E\begin{bmatrix}\epsilon_1 & \epsilon_2 & \dots & \epsilon_n\end{bmatrix}\begin{bmatrix}\epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n\end{bmatrix}$$

$$= E\begin{bmatrix}\epsilon_1^2 & \epsilon_1\epsilon_2 & \dots & \epsilon_1\epsilon_n \\ \epsilon_2\epsilon_1 & \epsilon_2^2 & \dots & \epsilon_2\epsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_n\epsilon_1 & \epsilon_n\epsilon_2 & \dots & \epsilon_n^2\end{bmatrix} \tag{9}$$

$$= \begin{bmatrix}E[\epsilon_1^2] & E[\epsilon_1\epsilon_2] & \dots & E[\epsilon_1\epsilon_n] \\ E[\epsilon_2\epsilon_1] & E[\epsilon_2^2] & \dots & E[\epsilon_2\epsilon_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[\epsilon_n\epsilon_1] & E[\epsilon_n\epsilon_2] & \dots & E[\epsilon_n^2]\end{bmatrix}$$

Recall, the regression assumption that the errors $\epsilon_i's$ are Normally distributed with mean 0 and variance σ^2. Thus, $E(\epsilon_i^2) = Var(\epsilon_i) = \sigma^2$ for all i. Additionally, recall the regression assumption that the errors are uncorrelated, i.e. $E(\epsilon_i\epsilon_j) = Cov(\epsilon_i, \epsilon_j) = 0$ for all i, j. Using these assumptions, we can write (9) as

$$E[\epsilon\epsilon^T] = \begin{bmatrix}\sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2\end{bmatrix} = \sigma^2\mathbf{I} \tag{10}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix.

Next, we show that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\epsilon$.

Recall that the $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ and $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Then,

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\end{aligned} \tag{11}$$

Using these two properties, we derive the form of the variance-covariance matrix for the coefficients. Note that the covariance matrix is $E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T]$

$$
\begin{aligned}
E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] &= E[(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} - \boldsymbol{\beta})(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon} - \boldsymbol{\beta})^T] \\
&= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}
\tag{12}
$$