

Linear Regression Explained

James Taylor

Suppose we have one quantitative response Y and p predictors X_1, X_2, \dots, X_p , and assume that there is an (approximately) linear relationship between them:

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Say we have training data consisting of n observations. We want to use these data to find approximations b_i for β_i ; we will then be able to predict future values of Y given $(X_1, \dots, X_p) = (x_1, \dots, x_p)$ via the formula

$$\hat{y} = b_0 + b_1 x_1 + \dots + b_p x_p$$

Our estimates for the model coefficients are going to be the values which minimize the training RSS (or equivalently, whose corresponding hyperplane has the smallest sum of squared “y-distances” between itself and the data). Letting our n observations be (x_i, y_i) for $i = 1, \dots, n$ where $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, we want to minimize

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_{i1} + \dots + b_p x_{ip}))^2$$

We minimize the RSS by differentiating with respect to b_j , setting this expression equal to zero, and simplifying the resulting equation to be of the form $\alpha_j \cdot b = \gamma_j$. We do this for each j , and then solve the following matrix equation for b :

$$\alpha b = \gamma$$

where α is the $(p+1) \times (p+1)$ matrix whose j th row is α_{j-1} , and γ is the $(p+1)$ -vector with j th entry γ_{j-1} .

More explicitly, for $j = 0$ we have

$$\begin{aligned} \frac{\partial}{\partial b_0}(RSS) &= \sum_{i=1}^n -2(y_i - (b_0 + b_1 x_{i1} + \dots + b_p x_{ip})) = 0 \\ \implies \sum_{i=1}^n (y_i - (b_0 + b_1 x_{i1} + \dots + b_p x_{ip})) &= 0 \\ \implies nb_0 + b_1 \sum_{i=1}^n x_{i1} + \dots + b_p \sum_{i=1}^n x_{ip} &= \sum_{i=1}^n y_i \end{aligned}$$

and for $j = 1, \dots, p$ we have

$$\begin{aligned} \frac{\partial}{\partial b_j}(RSS) &= \sum_{i=1}^n -2x_{ij}(y_i - (b_0 + b_1 x_{i1} + \dots + b_p x_{ip})) = 0 \\ \implies \sum_{i=1}^n x_{ij}(y_i - (b_0 + b_1 x_{i1} + \dots + b_p x_{ip})) &= 0 \\ \implies \sum_{i=1}^n x_{ij}(b_0 + b_1 x_{i1} + \dots + b_p x_{ip}) &= \sum_{i=1}^n x_{ij} y_i \\ \implies b_0 \sum_{i=1}^n x_{ij} + b_1 \sum_{i=1}^n x_{ij} x_{i1} + \dots + b_p \sum_{i=1}^n x_{ij} x_{ip} &= \sum_{i=1}^n x_{ij} y_i \end{aligned}$$

These equations give rise to the following matrix equation:

$$\begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ip} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{pmatrix}$$

Let x_{*j} be the n -vector consisting of all n observations of the j th predictor X_j . If we also set x_{*0} to be the n -vector containing all 1's, then this equation becomes

$$\begin{pmatrix} x_{*0} \cdot x_{*0} & x_{*0} \cdot x_{*1} & x_{*0} \cdot x_{*2} & \cdots & x_{*0} \cdot x_{*p} \\ x_{*1} \cdot x_{*0} & x_{*1} \cdot x_{*1} & x_{*1} \cdot x_{*2} & \cdots & x_{*1} \cdot x_{*p} \\ x_{*2} \cdot x_{*0} & x_{*2} \cdot x_{*1} & x_{*2} \cdot x_{*2} & \cdots & x_{*2} \cdot x_{*p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{*p} \cdot x_{*0} & x_{*p} \cdot x_{*1} & x_{*p} \cdot x_{*2} & \cdots & x_{*p} \cdot x_{*p} \end{pmatrix} b = \begin{pmatrix} x_{*0} \cdot y \\ x_{*1} \cdot y \\ x_{*2} \cdot y \\ \vdots \\ x_{*p} \cdot y \end{pmatrix}$$

Notice that the first matrix is just $X^T X$, and the RHS is $X^T y$, where X is the $n \times (p+1)$ matrix

$$\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

So our matrix equation is now

$$X^T X b = X^T y$$

and thus

$$b = (X^T X)^{-1} X^T y$$

However, depending on how large X is, inverting $X^T X$ could be extremely resource intensive. A way around this is to compute the QR decomposition of $X^T X$; that is, we find an orthogonal matrix Q and an upper triangular matrix R such that $X^T X = QR$. (Note: A QR decomposition exists for every matrix.) We'll still need to solve $QRb = X^T y$, but since Q is orthogonal we have $Q^{-1} = Q^T$, and thus we only need to solve

$$b = R^{-1} Q^T X^T y = R^{-1} (XQ)^T y$$

So now the only matrix we need to invert is the triangular matrix R , which is a much simpler task.