The Noble Eightfold Path to Linear Regression

Asher Labovich

January 2025

1 Introduction

Denote x_i the vector of p measurements, y_i the target response. N data points are then denoted as $(x_i, y_i)_{i=1}^N$.

We want to approximate $y_i = \phi_0 + \sum_{j=1}^p \phi_j x_{ij} + e_i$, for all i. e_i represents the residual for y_i. We want the vector Φ of coefficients so that $y = X\Phi + e$, where X has a vector of 1s at the left.

The goal is to find Φ that minimizes $||e||_2^2|| = e^T e$. Thus, we want to minimize

$$(y - X\Phi)^T(y - X\Phi) = \sum_{i=1}^N (y_i - \phi_0 - \sum_{j=1}^p \phi_j x_{ij})^2$$

2 Solutions

2.1 Partial Derivatives

We take the partial derivative of each ϕ_j. Thus, we have that

$$\begin{aligned}
\frac{\partial ||e||}{\phi_0} &= -2\sum_{i=1}^N (y_i - \phi_0 - \sum_{j=1}^p \phi_j x_{ij}) &= 0 \\
\frac{\partial ||e||}{\phi_1} &= -2\sum_{i=1}^N x_{i1}(y_i - \phi_0 - \sum_{j=1}^p \phi_j x_{ij}) &= 0 \\
\frac{\partial ||e||}{\phi_{k\neq 0}} &= -2\sum_{i=1}^N x_{ik}(y_i - \phi_0 - \sum_{j=1}^p \phi_j x_{ij}) &= 0
\end{aligned}$$

We can simplify these equations to get that

$$\sum_{i=1}^{N} y_i = \phi_0 \sum_{i=1}^{N} 1 + \sum_{i=1}^{N} \sum_{j=1}^{p} \phi_j x_{ij}$$
$$\sum_{i=1}^{N} y_i x_{ik} = \phi_0 \sum_{i=1}^{N} x_{ik} + x_{ik} \sum_{i=1}^{N} \sum_{j=1}^{p} \phi_j x_{ij}$$

Representing them a little nicer, we find that

$$\sum_{i=1}^{N} y_i = \phi_0 \sum_{i=1}^{N} 1 + \phi_1 \sum_{i=1}^{N} x_{i1} + \phi_2 \sum_{i=1}^{N} x_{i2} + \ldots + \phi_p \sum_{i=1}^{N} x_{ip}$$
$$\sum_{i=1}^{N} y_i x_{i1} = \phi_0 \sum_{i=1}^{N} x_{i1} + \phi_1 \sum_{i=1}^{N} x_{i1} x_{i1} + \phi_2 \sum_{i=1}^{N} x_{i2} x_{i1} + \ldots + \phi_p \sum_{i=1}^{N} x_{ip} x_{i1}$$
$$\vdots\vdots$$
$$\sum_{i=1}^{N} y_i x_{ip} = \phi_0 \sum_{i=1}^{N} x_{ip} + \phi_1 \sum_{i=1}^{N} x_{i1} x_{ip} + \phi_2 \sum_{i=1}^{N} x_{i2} x_{ip} + \ldots + \phi_p \sum_{i=1}^{N} x_{ip} x_{ip}$$

We recognize that the left side is equal to $X^T y$. The right side is equal to $X^T X \Phi$. If $(X^T X)^{-1}$ exists, then we have that $\Phi = (X^T X)^{-1} X^T y$.

We note that $X^T y = X^T X \Phi \implies X^T(y - X\Phi) = X^T e = 0$. Thus, $\sum_{i=1}^{N} e_i = 0, \sum_{i=1}^{N} e_i x_{ip} = 0$ for all p. Quite cool! Regardless of the form of $y = f(x)$, the linear appoximation will always have residuals sum to zero, as well as weighted sum of residuals for any one variable over N sum to 0.

2.2 Matrix Calculus

We have that $||e|| = (y - X\Phi)^T(y - X\Phi) = y^T y - (X\Phi)^T y - y^T(X\Phi) + \Phi^T X^T X \Phi$. Taking the vector derivative with respect to Φ, we get that $\frac{\partial ||e||}{\partial \mathbf{\Phi}} = 0 - 2X^T y + 2X^T X \Phi = 0$. Thus, we have that $\Phi = (X^T X)^{-1} X^T y$, should the inverse exist.

2.3 Pseudoinverse

All matrices have a "pseudoinverse", a matrix X^+ fulfilling the following requirements:

1. $X^+ X$ and $X X^+$ are symmetric.
2. $X X^+ X$ = X
3. $X^+ X X^+ = X^+$

We can reduce the equation for $||e||$ so that it equals

$$
\begin{aligned}
(y - X\Phi)^T(y - X\Phi) = \\
(XX^+y - XX^+y + y - X\Phi)^T(y - X\Phi) = \\
(XX^+y - X\Phi)^T(y - X\Phi) + y^T(I - XX^+)^T(y - X\Phi) = \\
(XX^+y - X\Phi)^T(y - X\Phi) + y^T(I - XX^+)^Ty - y^T(I - XX^+)^T(X\Phi) = \\
(XX^+y - X\Phi)^T(y - X\Phi) + y^T(I - XX^+)^Ty - y^T(X - XX^+X)(\Phi) = \\
(XX^+y - X\Phi)^T(y - X\Phi) + y^T(I - XX^+)^Ty - y^T(X - X)(\Phi) = \\
(XX^+y - X\Phi)^T(y - X\Phi) + y^T(I - XX^+)y
\end{aligned}
$$

Since the second term is constant with respect to Φ, we only care about minimizing the first term. Thus,

$$
\begin{aligned}
(XX^+y - X\Phi)^T(y - X\Phi) = \\
(X^+y - \Phi)^TX^T(y - X\Phi) = \\
(X^+y - \Phi)^T((XX^+X)^Ty - X^TX\Phi) = \\
(X^+y - \Phi)^T(X^TXX^+y - X^TX\Phi) = \\
(X^+y - \Phi)^TX^TX(X^+y - \Phi) = \\
||X(X^+y - \Phi)|| = \\
||X||||X^+y - \Phi||
\end{aligned}
$$

This is minimized when $\Phi = X^+y$. When $(X^TX)^{-1}$ exists, then $X^+ = (X^TX)^{-1}X^T$. However, this is quite a bit broader than the previous solution (and something I didn't know when starting this! Of course, this doesn't solve the problem of there still being infinite solns when X^TX is singular)

2.4 Statistical Approach

I skip this section since it is better covered in the next.

2.5 Normal Projection Approach

If we think of the columns of X as vectors in an (n+1)-dimensional space, we want to project y onto the column space of X with as little error as possible. Φ gives us this lowest error. This error must be orthogonal to the column space of X, or else there would be a better projection.

So, each column of X must have an inner product of 0 with e. So, $X^Te = 0$. We thus know that

$$0 = X^Te = X^T(y - X\Phi) = X^Ty - X^TX\Phi \implies \Phi = (X^TX)^{-1}X^Ty$$

2.6 Physics Approach

We can think of each point as the end of a spring connected to the line underneath them, in (n+1)-dimensional space. (Important: they cannot be slanted, they must be parallel to the "n+1"th dimension, or the dimension with y-coordinates). Then, the force of each spring is proportional to e_i, the distance of the spring to the line. For the line to be in equilibrium, we have that the forces must sum to 0 in every dimension. So, $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} e_i x_{ij} = 0$. Therefore, $X^T e = 0$, so

$$e = y - X\Phi \implies X^T e = X^T y - X^T X\Phi \implies \Phi = (X^T X)^{-1} X^T y$$