Least Squares Linear Regression

Consider again a linear regression model with $p$ inputs:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \text{noise}$$

The least squares estimates for the parameters, based on training data $(x_1, y_1), \ldots, (x_N, y_N)$, are those that minimize

$$\text{RSS}(\beta) = \sum_{i=1}^{N} \left[ y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \right]^2$$

Let’s put the inputs into a matrix, $X$, along with a column of 1’s, so that $X_{i1} = 1$ and $X_{ij} = x_{i,j-1}$ (input $j-1$ for the $i$’th training case). Also, put the responses into a column vector, $y$. Letting $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ be another column vector, we can then write

$$\text{RSS}(\beta) = |y - X\beta|^2 = (y - X\beta)^T(y - X\beta)$$

We denote the least squares estimates by $\hat{\beta}$, and the “fitted” response values in the training cases by $\hat{y} = X\hat{\beta}$. Note that $y - X\beta = y - \hat{y}$ are the residuals.
Finding the Least Squares Estimates

The minimum of $\text{RSS}(\beta)$ will occur where its gradient (vector of partial derivatives) is zero. This gradient vector is

$$\frac{\partial}{\partial \beta} \left[ (y - X\beta)^T(y - X\beta) \right] = -2X^T(y - X\beta)$$

Setting this to zero, we get $X^Ty = X^TX\beta$. Solving for $\beta$ gives

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

This can be computed as above, or better using the Cholesky decomposition of $X^TX$, or using an orthogonalization procedure on $X$.

This assumes that $X^TX$ is non-singular, so that there is a unique solution. When $p$ is greater than $N$, this will not be the case!