# 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

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

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

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

We denote the least squares estimates by $\hat{\beta}$, and the "fitted" response values in the training cases by $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$. Note that $\mathbf{y} - \mathbf{X}\beta \;=\; \mathbf{y} - \hat{\mathbf{y}}$ 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

$$\frac{\partial}{\partial\beta}\Big[(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\Big] \;=\; -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

Setting this to zero, we get $\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\beta$. Solving for $\beta$ gives

$$\hat{\beta} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

This can be computed as above, or better using the Cholesky decomposition of $\mathbf{X}^T\mathbf{X}$, or using an orthogonalization procedure on $\mathbf{X}$.

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