Ordinary Least Squares (Linear Least Squares)

A model that is linear in its coefficients, $\theta_j$, is given by $y(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + ....$ For a model with $M$ linear coefficients this can be written as

$$y(x) = \sum_{j=1}^{M} \theta_j f_j(x)$$

At $x_i$ this becomes

$$y(x_i) = \sum_{j=1}^{M} \theta_j f_j(x_i)$$

There are $N$ data points corresponding to observations made at $x_1 ... x_N$, so lets define a 'design matrix' $\mathbf{X}$ of size $N \times M$ such that $X_{i,j} = f_j(x_i)$. Now

$$y(x_i) = \sum_{j=1}^{M} \theta_j X_{i,j}$$

If the data is normally distributed with a known variance of $\sigma^2$ then the Likelihood is

$$L(\theta) = P(\vec{y}|\theta) = \prod_{i=1}^{N} \frac{1}{1/\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \sum_{j=1}^{M} \theta_j X_{i,j})^2}$$

Finding maximum of this expression (by varying $\theta$) is the same as finding the minimum of the negative, which is equivalent to finding the minimum of the negative of the log of the expression. So look for the minimum of:

$$S(\theta) \equiv \sum_{i=1}^{N} (y_i - \sum_{j}^{M} \theta_j X_{i,j})^2$$

Note that constant terms are dropped since we only care about where the minimum is found, not its value. The minimum is found when for each $k$,

$$\frac{\partial S}{\partial \theta_k} = 0$$

which written out is:

$$\frac{\partial S}{\partial \theta_k} = \sum_{i=1}^{N} (y_i - \sum_{j=1}^{M} \theta_j X_{i,j}) X_{i,k} = 0$$

Rearranging gives:

$$\sum_{i=1}^{N} y_i X_{i,k} = \sum_{i=1}^{N} \sum_{j=1}^{M} \theta_j X_{i,j} X_{i,k}$$

Since $\mathbf{y}$ is $N \times 1$, and $\mathbf{X}$ is $N \times M$, the left side looks like ($k$th row of $\mathbf{X}'$) $\times$ $\mathbf{y}$. Since $\theta$ is $M \times 1$, the right side looks like ($k$th row of $\mathbf{X}'$)$\times$ ( ($j$th row of $\mathbf{X}$) $\times$ $\theta$).

After the summations, the only remaining index is the coefficient index. We can write this so it relates two vectors of size $M \times 1$ by writing:

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\theta$$

And finally, to solve for $\theta$,

$$\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

Note the 'hat' above $\theta$ means the estimated $\theta$.