Least Squares Derivation
The standard linear model, independent variables $\hat{B}$, depedent variable $\hat{Y}$:

$$\hat{Y} = \hat{B_0} + \sum_{j=1}^{p} X_j \hat{B}_j.$$

The shorthand where we add a column of $\vec{1}$ to X and include $\hat{B_0}$ in the vector of coefficients $\hat{B}$

$$\hat{Y} = X^T \hat{B}.$$

Denote least-squares as RSS:

$$RSS(B) = \sum_{i=1}^{N} (y_i - x_i^T B)^2.$$

Recall from optimization that a linear equation is always convex AND concave, so a global minimum must exist.
We can rewrite the Least-Squares Equation in terms of matrices:

$$RSS(B) = (y - XB)^T (y - XB).$$

$$= y^T y - y^T XB - y(XB)^T + (XB)^T (XB) y^T y \text{ is a scalar.}$$

Differentiating w.r.t. B:

$$\frac{\partial RSS(B)}{\partial B} = 0 + -2X^T y + 2X^T XB = 0.$$

$$X^T y - X^T XB = 0.$$

$$X^T y = X^T XB.$$

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

■