Least Squared Estimators

Let the linear regression be $Y = X\beta + \varepsilon$, where the dimensions are $Y(N \times 1)$, $X(N \times k)$, $\beta(k \times 1)$, and $\varepsilon(N \times 1)$. A simplifying assumption is that $\varepsilon \sim N(0, \sigma^2 I)$

First, the LSE (Least Squared Estimators) is

$$\hat{\beta}_{LSE} = (X^T X)^{-1} X^T Y$$

Proof (in matrix format):

$$\hat{\beta}_{LSE} = \mathrm{argmin} \|Y - X\beta\|_2^2$$

$$\begin{aligned} \|Y - X\beta\|_2^2 &= (Y - X\beta)^T (Y - X\beta) = (Y^T - \beta^T X^T)(Y - X\beta) \\ &= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta \end{aligned}$$

(Note that $Y^T X\beta = \beta^T X^T Y$ is a scalar.)

Taking the gradient with respect to β

$$\frac{\partial \|Y - X\beta\|_2^2}{\partial \beta} = \frac{\partial (Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta)}{\partial \beta} = -2X^T Y + (X^T X + X^T X)\beta = 2X^T X\beta - 2X^T Y$$

The equation above used the following two rules in matrix calculus

$$\frac{\partial x^T A}{\partial x} = A$$

$$\frac{\partial x^T A x}{\partial x} = (A + A^T)x$$

where x is a $(k \times 1)$vector and A is a $(k \times n)$ matrix.

Setting the gradient to zero yields

$$2X^T X\beta - 2X^T Y = 0$$

$$\downarrow$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$