A Quick Return to Vectors and Matrices

In the simple linear regression setting:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\underset{\sim}{a}'\underset{\sim}{b} = (a_1 \dots a_n)\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \sum_{i=1}^{n} a_i b_i$$

Vectors: $\underset{\sim}{a}_{n\times 1} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$, $\underset{\sim}{b}_{n\times 1} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$

length n

$$\underset{2\times n}{C'}\ \underset{\sim\, n\times 1}{b} = \begin{pmatrix} c_{11} \dots c_{n1} \\ c_{12} \dots c_{n2} \end{pmatrix}\begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} \underset{\sim}{c}_1' b \\ \underset{\sim}{c}_2' b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} c_{i1} b_i \\ \sum_{i=1}^{n} c_{i2} b_i \end{pmatrix}$$

Matrices: $C_{n\times 2} = \begin{pmatrix} \underset{\sim}{c}_{1\, n\times 1} & \underset{\sim}{c}_{2\, n\times 1} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ \vdots & \vdots \\ c_{n1} & c_{n2} \end{pmatrix}$

How Vectors and Matrices Make Multiple Linear Regression Easier...

$$\underset{\sim\, p\times 1}{b} = (X'X)^{-1} X'\underset{\sim}{y}$$

$$\underset{n\times p}{X} = \begin{pmatrix} \underset{\sim}{x}_0 & \cdots & \underset{\sim}{x}_{p-1} \end{pmatrix}$$

$x_{i0} = 1$ for all i

$$\bar{x} = \underbrace{\begin{pmatrix} 1 & \cdots & 1 \end{pmatrix}}_{\underset{\sim}{1}_n'} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \times \frac{1}{n}$$

$$\bar{x} = (\underset{\sim}{1}_n' \underset{\sim}{1}_n)^{-1} \underset{\sim}{1}_n' \underset{\sim}{x}$$

$$X'X = \begin{pmatrix} \underset{\sim}{x}_0' \underset{\sim}{x}_0 & \underset{\sim}{x}_0' \underset{\sim}{x}_1 & \cdots & \underset{\sim}{x}_0' \underset{\sim}{x}_{p-1} \\ \underset{\sim}{x}_0' \underset{\sim}{x}_1 & \underset{\sim}{x}_1' \underset{\sim}{x}_1 & \cdots & \underset{\sim}{x}_1' \underset{\sim}{x}_{p-1} \\ \vdots & & \ddots & \vdots \\ \underset{\sim}{x}_0' \underset{\sim}{x}_{p-1} & \cdots & & \underset{\sim}{x}_{p-1}' \underset{\sim}{x}_{p-1} \end{pmatrix}$$

Simple Linear Regression in Matrix and Vector Notation

$$\underset{n\times 2}{X} = \begin{pmatrix} \underset{\sim}{1}_n & \underset{\sim}{x} \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \qquad b = (X'X)^{-1}X'\underset{\sim}{y}$$

$$X'X = \begin{pmatrix} \underset{\sim}{1}_n'\underset{\sim}{1}_n & \underset{\sim}{1}_n'\underset{\sim}{x} \\ \underset{\sim}{1}_n'\underset{\sim}{x} & \underset{\sim}{x}'\underset{\sim}{x} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)(\sum x_i)} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

Simple linear Regression in Matrix and Vector Notation

$$b = (X'X)^{-1} X' \underset{\sim}{y}$$

$$= \frac{1}{n\sum x_i^2 - (\sum x_i)(\sum x_i)} \left(\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \underset{\sim}{1}_n{}' \underset{\sim}{y} \\ \underset{\sim}{x}' \underset{\sim}{y} \end{pmatrix} \right)$$

$$= \frac{1}{n\sum x_i^2 - (\sum x_i)(\sum x_i)} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

$$= \frac{1}{n\sum x_i^2 - (\sum x_i)(\sum x_i)} \begin{pmatrix} (\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum y_i x_i) \\ (-\sum x_i)(\sum y_i) + n\sum x_i y_i \end{pmatrix}$$

$$\underset{\sim}{b} = \frac{1}{n\sum x_i^2 - (\sum x_i)(\sum x_i)} \begin{pmatrix} (\sum x_i^2)(\sum y_i) - (\sum x_i)(\sum y_i x_i) \\ (-\sum x_i)(\sum y_i) + n\sum x_i y_i \end{pmatrix}$$

Easy to see the relationship to our previous definitian

$$\text{if } \bar{x} = 0 \Rightarrow \sum x_i = 0$$

$$\underset{\sim}{b} = \frac{1}{n\sum x_i^2} \begin{pmatrix} \sum x_i^2 \sum y_i \\ n\sum x_i y_i \end{pmatrix} = \begin{pmatrix} \bar{Y} \\ \frac{\sum x_i y_i}{\sum x_i^2} \end{pmatrix}$$

This looks like our earlier definition!

Multiple Linear Regression Model with Normal Errors

$\underset{n\times 1}{\underset{\sim}{y}} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ equivalent to $y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i$] includes p-1 covariates

y: response vector; X: nxp covariate matrix, first column is a vector of 1's

* $E\{\underset{\sim}{\varepsilon}\} = \underset{\sim}{0}$ equivalent to $E\{\varepsilon_i\} = 0$
* $\sigma^2\{\underset{\sim}{\varepsilon}\} = \sigma^2 I_n$ equivalent to $\sigma^2\{\varepsilon_i\} = \sigma^2$, $\sigma\{\varepsilon_i, \varepsilon_j\} = 0$, $i \neq j$
* $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

$\Rightarrow E\{\underset{\sim}{y}\} = X\underset{\sim}{\beta}$, $\sigma^2\{\underset{\sim}{y}\} = \sigma^2 I_n$

Coefficient of Multiple Determination
$R^2 = SSR/SSTO = 1 - SSE/SSTO$

* what proportion of the variability in y is explained by the covariates
* R^2 is nondecreasing as you add more

We have estimators:

* $\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{y}$
* $\underset{\sim}{\hat{Y}} = X\underset{\sim}{b}$
* $\underset{\sim}{e} = Y - \hat{Y}$
* $s^2 = \dfrac{\underset{\sim}{e}'\underset{\sim}{e}}{n-p}$

ANOVA sums-of-squares

$SSE = \underset{\sim}{e}'\underset{\sim}{e}$ has n-p degrees freedom

$SSTO$ (n-1 df) $= \sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - \frac{1}{n}\underset{\sim}{1}_n' y)'(y - \frac{1}{n}\underset{\sim}{1}_n' y)$

SSR (p-1 df.) $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = (\hat{y} - \frac{1}{n}\underset{\sim}{1}_n' y)'(\hat{y} - \frac{1}{n}\underset{\sim}{1}_n' y)$

$MSE = \underset{\sim}{e}'\underset{\sim}{e}/(n-p)$ $\quad MSTO = SSTO/n-1$ $\quad MSR = SSR/p-1$

Coefficient of Multiple Determination

$$R^2 = SSR/SSTO = 1 - SSE/SSTO$$

* what proportion of the variability in y is explained by the covariates
* R^2 is nondecreasing as you add more covariates

$\Longrightarrow$ If you decide to pick your covariates to maximize R^2, you will always decide to use all of them!

Brings us to...

Adjusted Coefficient of Multiple Determination

$$R_a^2 = 1 - \frac{MSE}{MSTO} = 1 - \frac{SSE/n-p}{SSTO/n-1}$$

* R_a^2 can decrease as more covariates are added

$\Longrightarrow$ If you were to pick your "best" model using R_a^2, it won't necessarily include all of your covariates

Inference about Regression Parameters

Under our multiple linear regression model,

$$\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{y}$$

* are unbiased $E\{\underset{\sim}{b}\} = \underset{\sim}{\beta}$ (equivalent to $\begin{pmatrix} E\{b_0\} \\ E\{b_1\} \\ \vdots \\ E\{b_{p-1}\} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$)

* $$\sigma^2\{\underset{\sim}{b}\} = \begin{pmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} & \cdots & \sigma\{b_0, b_{p-1}\} \\ \sigma\{b_0, b_1\} & \sigma^2\{b_1\} & \cdots & \sigma\{b_1, b_{p-1}\} \\ \vdots & \vdots & & \vdots \\ \sigma\{b_0, b_{p-1}\} & \sigma\{b_1, b_{p-1}\} & \cdots & \sigma^2\{b_{p-1}\} \end{pmatrix}$$

$$= \sigma^2 (X'X)^{-1}$$

* can get an estimate of $\sigma^2\{\underset{\sim}{b}\}$, $s^2\{\underset{\sim}{b}\} = s^2(X'X)^{-1}$