Multiple Regression

Dr. Frank Wood

Review: Matrix Regression Estimation

- We can solve this equation

$$\underset{2\times 2}{\mathbf{X'X}}\ \underset{2\times 1}{\mathbf{b}} = \underset{2\times 1}{\mathbf{X'Y}}$$

(if the inverse of X'X exists) by the following

$$(\mathbf{X'X})^{-1}\mathbf{X'Xb} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

and since $(\mathbf{X'X})^{-1}\mathbf{X'X} = \mathbf{I}$

we have $\underset{2\times 1}{\mathbf{b}} = \underset{2\times 2}{(\mathbf{X'X})^{-1}}\ \underset{2\times 1}{\mathbf{X'Y}}$

Least Squares Solution

- The matrix normal equations can be derived directly from the minimization of

$$Q = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

w.r.t. to β

Do this on board.

Fitted Values and Residuals

- Let the vector of the fitted values be

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

in matrix notation we then have

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \underset{n\times 2}{\mathbf{X}} \underset{2\times 1}{\mathbf{b}}$$

Hat Matrix – Puts hat on Y

- We can also directly express the fitted values in terms of only the X and Y matrices

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

and we can further define H, the "hat matrix"

$$\underset{n\times 1}{\hat{\mathbf{Y}}} = \underset{n\times n}{\mathbf{H}}\ \underset{n\times 1}{\mathbf{Y}} \qquad \underset{n\times n}{\mathbf{H}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- The hat matrix plans an important role in diagnostics for regression analysis.

write H on board

Hat Matrix Properties

- The hat matrix is symmetric
- The hat matrix is idempotent, i.e.

$$\mathbf{HH} = \mathbf{H}$$

demonstrate on board

Residuals

- The residuals, like the fitted values of \hat{Y_i} can be expressed as linear combinations of the response variable observations Y_i

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$\underset{n\times 1}{\mathbf{e}} = \underset{n\times 1}{\mathbf{Y}} - \underset{n\times 1}{\hat{\mathbf{Y}}} = \underset{n\times 1}{\mathbf{Y}} - \underset{n\times 1}{\mathbf{Xb}}$$

$$\underset{n\times 1}{\mathbf{e}} = \underset{n\times n}{(\mathbf{I}} - \underset{n\times n}{\mathbf{H})} \underset{n\times 1}{\mathbf{Y}}$$

Covariance of Residuals

- Starting with $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
  we see that $\boldsymbol{\sigma}^2\{\mathbf{e}\} = (\mathbf{I} - \mathbf{H})\boldsymbol{\sigma}^2\{\mathbf{Y}\}(\mathbf{I} - \mathbf{H})'$
  but $\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\} = \sigma^2\mathbf{I}$
  which means that

$$\boldsymbol{\sigma}^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{I}(\mathbf{I} - \mathbf{H})$$

$$= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})$$

and since I-H is idempotent (check) we have

$$\underset{n\times n}{\boldsymbol{\sigma}^2\{\mathbf{e}\}} = \sigma^2(\mathbf{I} - \mathbf{H})$$

we can plug in MSE for σ^2 as an estimate

ANOVA

- We can express the ANOVA results in matrix form as well, starting with

$$SSTO = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$$

where

$$\mathbf{Y}'\mathbf{Y} = \sum Y_i^2 \qquad \frac{(\sum Y_i)^2}{n} = \left(\frac{1}{n}\right) \mathbf{Y}'\mathbf{J}\mathbf{Y}$$

J is matrix of all ones, do 3x3 example

leaving

$$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right) \mathbf{Y}'\mathbf{J}\mathbf{Y}$$

SSE

- Remember

$$SSE = \sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

- We have

$$SSE = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y}'\mathbf{Y} - 2\mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}$$

derive this on board

$$SSE = \mathbf{Y}'\mathbf{Y} - 2\mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}}_{\mathbf{b}}$$

and this

$$= \mathbf{Y}'\mathbf{Y} - 2\mathbf{b}'\mathbf{X}'\mathbf{Y} + \mathbf{b}'\mathbf{I}\mathbf{X}'\mathbf{Y}$$

- Simplified

$$SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$$

SSR

- It can be shown that
  - for instance, remember SSR = SSTO-SSE

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

$$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} \qquad SSE = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$$

write these on board

Tests and Inference

- The ANOVA tests and inferences we can perform are the same as before
- Only the algebraic method of getting the quantities changes
- Matrix notation is a writing short-cut, not a computational shortcut

Quadratic Forms

- The ANOVA sums of squares can be shown to be quadratic forms. An example of a quadratic form is given by

$$5Y_1^2 + 6Y_1Y_2 + 4Y_2^2$$

- Note that this can be expressed in matrix notation as (where A is a symmetric matrix)

$$[Y_1 \quad Y_2]\begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \mathbf{Y}'\mathbf{A}\mathbf{Y}$$

do on board

Quadratic Forms

- In general, a quadratic form is defined by

$$\underset{1\times 1}{\mathbf{Y}'\mathbf{A}\mathbf{Y}} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} Y_i Y_j \qquad \text{where } a_{ij} = a_{ji}$$

  A is the matrix of the quadratic form.
- The ANOVA sums SSTO, SSE, and SSR are all quadratic forms.

ANOVA quadratic forms

- Consider the following re-expression of b'X'

$$\mathbf{b}'\mathbf{X}' = (\mathbf{X}\mathbf{b})' = \hat{\mathbf{Y}}' \qquad \mathbf{b}'\mathbf{X}' = (\mathbf{H}\mathbf{Y})'$$

$$\mathbf{b}'\mathbf{X}' = \mathbf{Y}'\mathbf{H}$$

- With this it is easy to see that

$$SSTO = \mathbf{Y}'\left[\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}$$

$$SSE = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$SSR = \mathbf{Y}'\left[\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}$$

Inference

- We can derive the sampling variance of the β vector estimator by remembering that

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{A}\mathbf{Y}$$

where A is a constant matrix

$$\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad \mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

which yields

$$\boldsymbol{\sigma}^2\{\mathbf{b}\} = \mathbf{A}\boldsymbol{\sigma}^2\{\mathbf{Y}\}\mathbf{A}'$$

Variance of b

- Since $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric we can write

$$\mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

and thus

$$\begin{aligned}\boldsymbol{\sigma}^2\{\mathbf{b}\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{I} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Variance of b

- Of course this assumes that we know σ^2. If we don't, we, as usual, replace it with the MSE.

$$\underset{2\times 2}{\sigma^2\{\mathbf{b}\}} = \begin{bmatrix} \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{\sum (X_i - \bar{X})^2} & \frac{-\bar{X}\sigma^2}{\sum (X_i - \bar{X})^2} \\ \frac{-\bar{X}\sigma^2}{\sum (X_i - \bar{X})^2} & \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \end{bmatrix}$$

$$\underset{2\times 2}{\mathbf{s}^2\{\mathbf{b}\}} = MSE(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{MSE}{n} + \frac{\bar{X}^2 MSE}{\sum (X_i - \bar{X})^2} & \frac{-\bar{X}MSE}{\sum (X_i - \bar{X})^2} \\ \frac{-\bar{X}MSE}{\sum (X_i - \bar{X})^2} & \frac{MSE}{\sum (X_i - \bar{X})^2} \end{bmatrix}$$

Mean Response

- To estimate the mean response we can create the following matrix

$$\underset{2\times 1}{\mathbf{X}_h} = \begin{bmatrix} 1 \\ X_h \end{bmatrix} \qquad \text{or} \qquad \underset{1\times 2}{\mathbf{X}'_h} = [1 \quad X_h]$$

- The fit (or prediction) is then

$$\hat{Y}_h = \mathbf{X}'_h \mathbf{b}$$

since

$$\mathbf{X}'_h \mathbf{b} = [1 \quad X_h] \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = [b_0 + b_1 X_h] = [\hat{Y}_h] = \hat{Y}_h$$

Variance of Mean Response

- Is given by

$$\sigma^2\{\hat{Y}_h\} = \sigma^2 \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$$

  and is arrived at in the same way as for the variance of β

- Similarly the estimated variance in matrix notation is given by

$$s^2\{\hat{Y}_h\} = MSE(\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

Wrap-Up

- Expectation and variance of random vector and matrices
- Simple linear regression in matrix form

- Next: multiple regression

Multiple Regression

- One of the most widely used tools in statistical analysis
- Matrix expressions for multiple regression are the same as for simple linear regression

Need for Several Predictor Variables

- Often the response is best understood as being a function of multiple input quantities
  - Examples
    - Spam filtering – regress the probability of an email being a spam message against thousands of input variables
    - Football prediction – regress the probability of a goal in some short time span against the current state of the game

First-Order Model with Two Predictor Variables

- When there are two predictor variables X_1 and X_2 the regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  is called a first-order model with two predictor variables.
- A first order model is linear in the predictor variables.
- X_{i1} and X_{i2} are the values of the two predictor variables in the ith trial

Functional Form

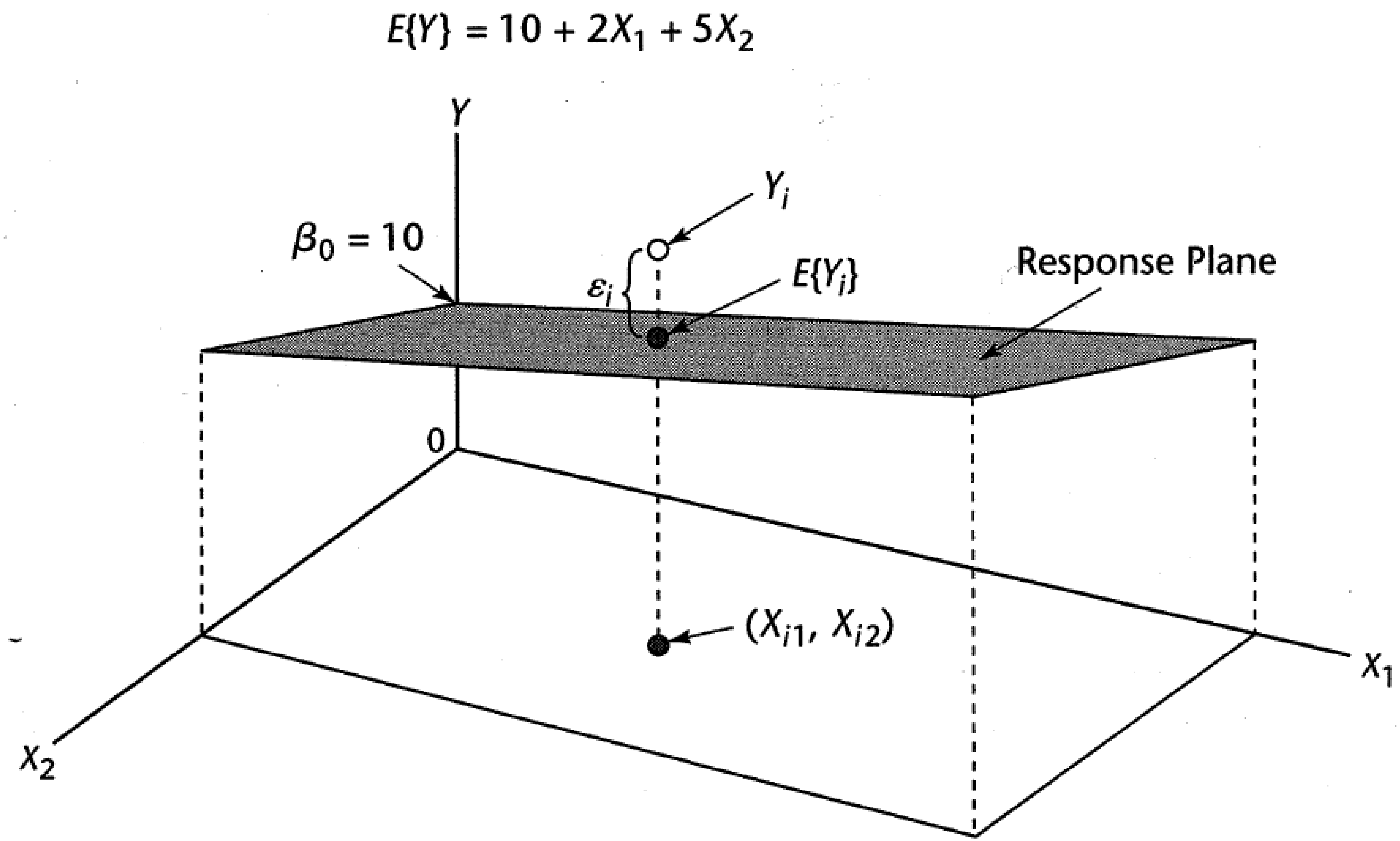
- Assuming noise equal to zero in expectation

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- The form of this regression function is of a plane
  - e.g.

$$E\{Y\} = 10 + 2X_1 + 5X_2$$

Regression (response) surface



Meaning of Regression Coefficients

- β_0 is the intercept when both X_1 and X_2 are zero
- β_1 indicates the change in the mean response E{Y} per unit increase in X_1 when X_2 is held constant
- β_2 – vice versa
- Example – fix $X_2 = 2$

$$E\{Y\} = 10 + 2X_1 + 5(2) = 20 + 2X_1 \qquad X_2 = 2$$

intercept changes but interpretation is clear

Terminology

- When the effect of X_1 on the mean response does not depend on the level of X_2 (and vice versa) the two predictor variables are said to have *additive effects* or *not to interact.*
- The parameters β_1 and β_2 are sometimes called *partial regression coefficients*

Comments

- A planar response surface may not always be appropriate, but even when not it is often a good approximate descriptor of the regression function in “local” regions of the input space
- The meaning of the parameters can be determined by taking partials of the regression function w.r.t. to each.

First order Model with >2 Predictor Variables

- Let there be p-1 predictor variables, then

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

which can also be written as

$$Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i$$

and if $X_{i0} = 1$ is also can be written as

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i \qquad \text{where } X_{i0} \equiv 1$$

Geometry of response surface

- In this setting the response surface is a hyperplane
- This is difficult to visualize but the same intuitions hold
  - Fixing all input variables, each β tells how much the response variable will grow or decrease according to its own (and only its own) input variable

General Linear Regression Model

- We have arrived at the general regression model. In general the $X_1, \ldots, X_{p-1}$ variables in the regression model do not have to represent different predictor variables, nor do they have to all be quantitative (continuous).
- The general model is

$$Y_i = \sum_{k=0}^{p-1} \beta_k X_{ik} + \varepsilon_i \qquad \text{where } X_{i0} \equiv 1$$

with response function (when $E\{\epsilon_i\} = 0$)

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

Qualitative (Discrete) Predictor Variables

- Until now we have (implicitly) focused on quantitative (continuous) predictor variables.
- Qualitative (discrete) predictor variables often arise in the real world
  - Examples
    - Patient sex: male/female/other
    - Goal scored in last minute: yes/no
    - Etc.

Example

- Regression model to predict the length of hospital stay (Y) based on the age (X_1) and gender (X_2) of the patient. Define X_2 as

$$X_2 = \begin{cases} 1 & \text{if patient female} \\ 0 & \text{if patient male} \end{cases}$$

- And use the standard first-order regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Example cont.

- Where

$$X_{i1} = \text{patient's age}$$
$$X_{i2} = \begin{cases} 1 & \text{if patient female} \\ 0 & \text{if patient male} \end{cases}$$

- If $X_2 = 0$ (i.e. patient is male) the response function is

$$E\{Y\} = \beta_0 + \beta_1 X_1$$

- otherwise it is

$$E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$$

- which is just another (parallel) linear response function with a different intercept

Polynomial Regression

- Polynomial regression models are special cases of the general regression model.
- They can contain squared and higher-order terms of the predictor variables.
- The response function becomes curvilinear.
- For example

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

which clearly has the same form as the general regression model.

General Regression

- Transformed variables
  - log Y, 1/Y
- Interaction effects

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Combinations

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i$$

- Key point – all linear in parameters!

General Regression Model in Matrix Terms

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \underset{n\times p}{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

General Linear Regression in Matrix Terms

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}} \underset{n\times p}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

- With $\mathbf{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$

  and

$$\underset{n\times n}{\boldsymbol{\sigma}^2\{\boldsymbol{\varepsilon}\}} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

- We have $\underset{n\times 1}{\mathbf{E}\{\mathbf{Y}\}} = \mathbf{X}\boldsymbol{\beta}$ and $\underset{n\times n}{\boldsymbol{\sigma}^2\{\mathbf{Y}\}} = \sigma^2\mathbf{I}$

Least Squares Estimation

- Same as before

$$\underset{2\times1}{\mathbf{b}} = \underset{2\times2}{(\mathbf{X}'\mathbf{X})^{-1}} \underset{2\times1}{(\mathbf{X}'\mathbf{X})}\, \mathbf{Y}$$

WRONG!!!!

- Maximum likelihood under iid normal error assumption results in same estimator
- Fitted values and residuals the same as before as well.

ANOVA

- The sums of squares derived before are the same here

$$SSTO = \mathbf{Y}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left[\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}$$

$$SSE = \mathbf{e}'\mathbf{e} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$$

$$SSR = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \left(\frac{1}{n}\right)\mathbf{Y}'\mathbf{J}\mathbf{Y} = \mathbf{Y}'\left[\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{Y}$$

but now we have to account for more parameters