SECTION 3

Matrix Notation, Properties of Least Squares Estimators

Matrix Notation
Simple Linear Regression

$$Y_1, \ldots Y_n \Rightarrow Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & X_1 \\ . & . \\ . & . \\ . & . \\ 1 & X_n \end{bmatrix}}_{\text{Design matrix}} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

Matrix Multiplication

$$\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix}$$

Matrix Notation

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathcal{E} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Matrix Notation

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1, \\
Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2, \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n
\end{aligned}
$$

- Can be written as:

$$\mathbf{Y} = \mathbf{X}\beta + \mathcal{E}$$

Matrix Notation

- An error covariance matrix shows the structure of covariances among different error terms. By definition, error covariance matrices are symmetric (the number in the i^{th} row and j^{th} column is the same as the number in the j^{th} row and i^{th} column). The number, or "element", in the i^{th} row and j^{th} column is the covariance between the i^{th} and j^{th} errors.

Error Covariance Matrix

$$\mathcal{E} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Var(\varepsilon) = \begin{bmatrix} \text{cov}(\varepsilon_1,\varepsilon_1) & \text{cov}(\varepsilon_1,\varepsilon_2) & \dots & \text{cov}(\varepsilon_1,\varepsilon_n) \\ \text{cov}(\varepsilon_1,\varepsilon_2) & \text{cov}(\varepsilon_2,\varepsilon_2) & & \\ & \vdots & \ddots & \\ \text{cov}(\varepsilon_1,\varepsilon_n) & & & \text{cov}(\varepsilon_n,\varepsilon_n) \end{bmatrix}$$

The errors have common variance and are uncorrelated :

$$\text{cov}(\varepsilon_i,\varepsilon_j) = \begin{cases} \sigma^2 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

So $Var(\varepsilon) = \sigma^2 I$ where I is the $n \times n$ identity matrix

Matrix Notation – Fitted Values and Residuals

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 X_1 \\ b_0 + b_1 X_2 \\ \vdots \\ b_0 + b_1 X_n \end{bmatrix} = \mathbf{X}\mathbf{b}$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Example – Section 1

x	y	$\hat{y}_i = 1 + 1.3x_i$	$r_i = y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	3	2.3	0.7	0.49
2	3	3.6	-0.6	0.36
4	7	6.2	0.8	0.64
5	6	7.5	-1.5	2.25
8	12	11.4	0.6	0.36

Fitted Values and Residuals

$$\hat{\mathbf{Y}} = \begin{bmatrix} 2.3 \\ 3.6 \\ 6.2 \\ 7.5 \\ 11.4 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 5 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 1 \\ 1.3 \end{bmatrix} = \begin{bmatrix} 1+1.3\times 1 \\ 1+1.3\times 2 \\ 1+1.3\times 4 \\ 1+1.3\times 5 \\ 1+1.3\times 8 \end{bmatrix}$$

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \begin{bmatrix} 0.7 \\ -0.6 \\ 0.8 \\ -1.5 \\ 0.6 \end{bmatrix}$$

Least Squares Estimates Matrix Notation

$$Define \quad b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

- The distance or squared error loss function

$$e'e = d(b) = (Y - Xb)^T (Y - Xb)$$

- Taking the derivative wrt b and setting to zero yields the normal equations

$$\frac{\partial d}{\partial b} = -2X^T(Y - Xb) = 0.$$

Least Squares Estimates Matrix Notation

- We can then see that:

$$X^T X b = X^T Y$$

$$b = (X^T X)^{-1} X^T Y.$$

Some important algebra

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \\ 1 & X_n \end{bmatrix}' \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{bmatrix}$$

Some important relationships and notation

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})x_i = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})Y_i = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_iY_i - n\overline{x}\overline{Y}$$

Using the results on the previous slides

$$X^T X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$$

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{then} \quad A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

$$\checkmark \quad (X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

Solving

$$b = (X^T X)^{-1} X^T Y = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{j=1}^n Y_j \\ \sum_{j=1}^n x_j Y_j \end{pmatrix}$$

$$= \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{j=1}^n Y_j - \sum_{i=1}^n x_i \sum_{j=1}^n x_j Y_j \\ n \sum_{j=1}^n x_j Y_j - \sum_{i=1}^n x_i \sum_{j=1}^n Y_j \end{pmatrix}$$

$$= \frac{1}{S_{xx}} \begin{pmatrix} \overline{Y} \sum_{i=1}^n x_i^2 - \overline{x} \sum_{j=1}^n x_j Y_j \\ \sum_{j=1}^n x_j Y_j - n\overline{x}\overline{Y} \end{pmatrix}$$

Solving

$$= \frac{1}{S_{xx}} \begin{pmatrix} \overline{Y}\sum_{i=1}^{n} x_i^2 - n\overline{Y}\overline{x}^2 + n\overline{Y}\overline{x}^2 - \overline{x}\sum_{j=1}^{n} x_j Y_j \\ \\ S_{xy} \end{pmatrix}$$

$$= \frac{1}{S_{xx}} \begin{pmatrix} \overline{Y} S_{xx} - \overline{x} S_{xy} \\ \\ S_{xy} \end{pmatrix}$$

$$= \begin{pmatrix} \overline{Y} - \overline{x}\frac{S_{xy}}{S_{xx}} \\ \\ \frac{S_{xy}}{S_{xx}} \end{pmatrix}.$$

The "hat matrix"

This form of the least-squares estimator leads to another useful matrix, since we may write

$$\hat{Y} = Xb = X(X^T X)^{-1} X^T Y = HY,$$

where $H = X(X^T X)^{-1} X^T$ is called the "hat matrix" since multiplying the response vector, Y, by H yields $\hat{Y}$. Similarly, the vector of residuals can be written as

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

Leverage

The diagonal elements of H are a measure of the *influence* of each data point. It turns out that the i^{th} diagonal element of H is:

$$h_{ii} = \frac{\sum_{j=1}^{n}(x_j - x_i)^2}{nS_{xx}},$$

and the value h_{ii} is referred to as the *leverage* of the i^{th} data point. The leverage h_{ii} quantifies how far away the i^{th} x value is from the rest of the x values. If the i^{th} x value is far away, the leverage h_{ii} will be large; and otherwise not.

We see that this is a measure of how far from the main body of the data each point is in the horizontal (or predictor) direction.

The leverage h_{ii} is a number between 0 and 1

Leverage

It turns out that the sum of all the leverages in a simple linear regression is equal to 2!:

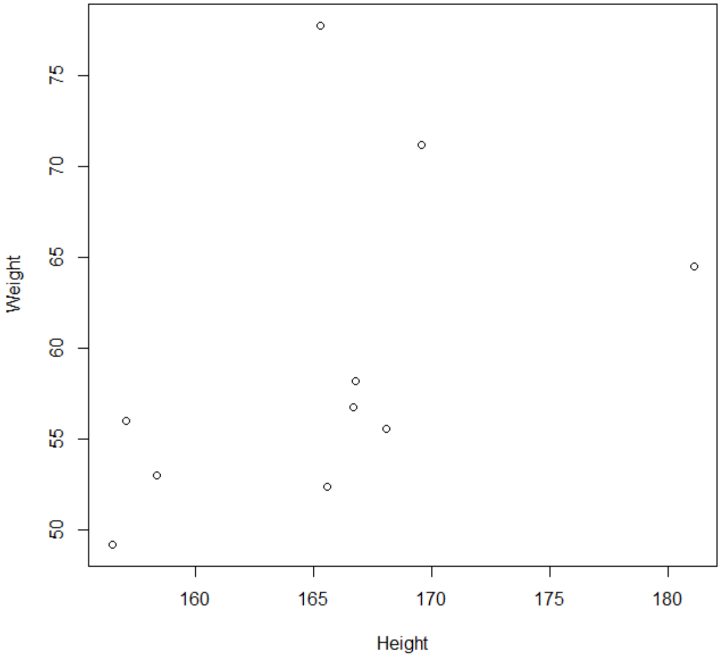
$$\sum_{i=1}^{n} h_{ii} = 2.$$

If all the data points had the same leverage, each of the h_{ii}'s should be equal to $2/n$.

So, any point which exceeds this value, and more particularly, any point which exceeds twice this value is a potentially influential point.

We will see this again later!

Leverage example worksheet2_women.csv on wattle



- Which observation do we think may have high leverage?

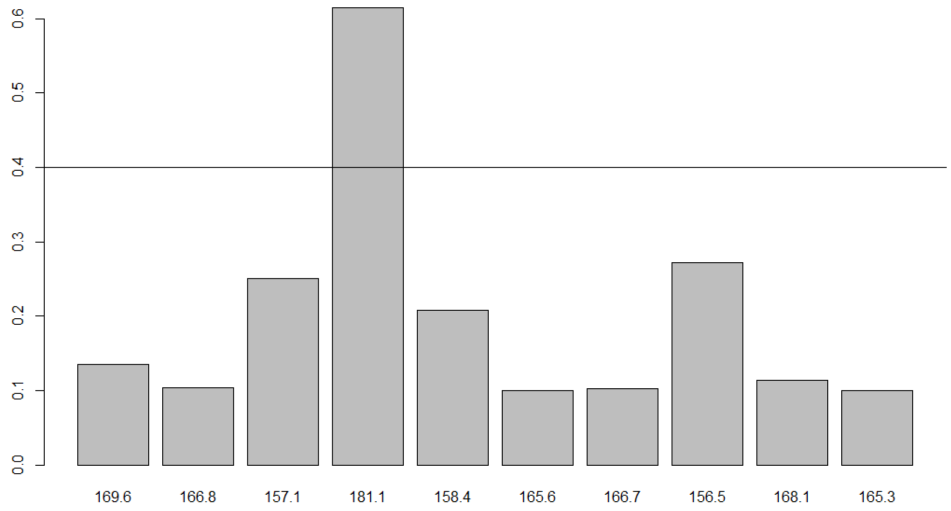
Example – plot of leverage for each point.

women<-read.csv("worksheet2_women.csv",header=F)

names(women)<-c("Height","Weight")

barplot(hat(Height),names.arg=Height)

abline(h=4/length(Height)



Leverage and Influence

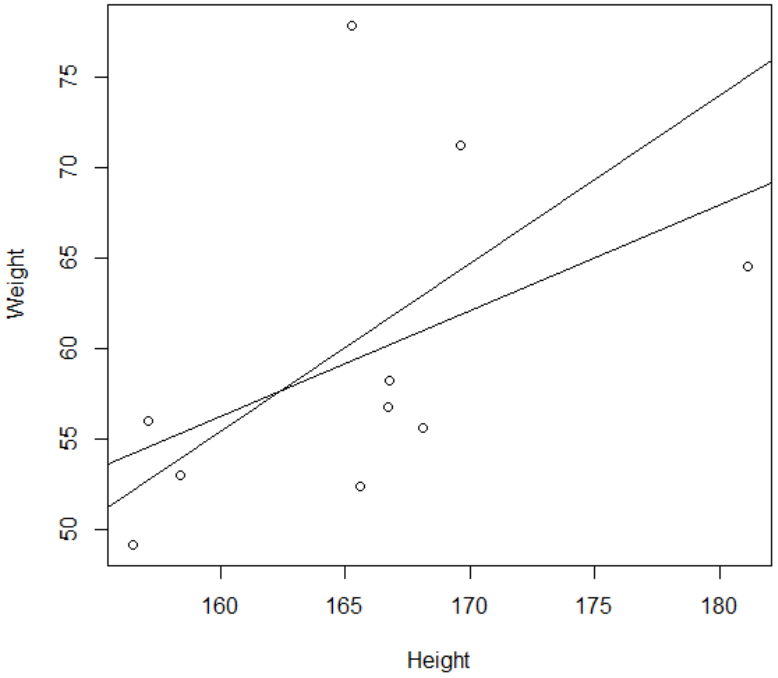
A point is highly influential if its removal from the dataset causes a dramatic change in the estimated parameters of the regression line.

One useful way of flagging the potentially influential data points is through the use of the *leverages*, h_{ii}

In order to see whether points with high leverage truly are influential, we examine how the fitted regression line would change once that flagged point is deleted from the data set.

```
>abline(lsfit(Height,Weight))

>abline(lsfit(Height[-4],Weight[-4]))
```



Properties of Least – Squares Estimators

The estimates b_0 and b_1 are unbiased:

$$E(b_1) = E(\frac{S_{xy}}{S_{xx}})$$

$$= E(\sum_{i=1}^{n} \frac{(x_i - \overline{x})(Y_i - \overline{Y})}{S_{xx}})$$

$$= E(\sum_{i=1}^{n} \frac{x_i Y_i - x_i \overline{Y} - \overline{x} Y_i + \overline{x}\,\overline{Y}}{S_{xx}})$$

$$= \sum_{i=1}^{n} \frac{(x_i - \overline{x}) E(Y_i)}{S_{xx}}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{S_{xx}}$$

$$= \frac{\beta_0 \sum_{i=1}^{n} (x_i - \overline{x}) + \beta_1 S_{xx}}{S_{xx}}$$

$$= \beta_1$$

x_i is an independent (or predictor) variable which is known exactly, while y is a dependent (or response) random variable.

Properties of Least-Squares Estimators

The estimates b_0 and b_1 are unbiased:

$$E(b_0) = E(\overline{Y} - b_1\overline{x}) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) - \beta_1\overline{x} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \beta_1\overline{x} = \beta_0.$$

Matrix Notation is easier!

since $Y = X\beta + \epsilon$ implies that the least-squares estimator, $b = (X^T X)^{-1} X^T Y$, may be written as:

$$(X^T X)^{-1} X^T (X\beta + \epsilon)$$

$$= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X^T \epsilon$$

$$= \beta + (X^T X)^{-1} X^T \epsilon,$$

which implies that $E(b) = E\{\beta + (X^T X)^{-1} X^T \epsilon\} = \beta + (X^T X)^{-1} X^T E(\epsilon) = \beta$, since $E(\epsilon) = 0$.

Variance – Matrix Notation

We know the least squares estimator $b = (X^TX)^{-1}X^TY$

$$Var(b) = Var\{\beta + (X^TX)^{-1}X^T\epsilon\}$$

employing the rule $Var(AZ) = AVar(Z)A^T$

$$= (X^TX)^{-1}X^TVar(\epsilon)\{(X^TX)^{-1}X^T\}^T$$

$$= (X^TX)^{-1}X^T(\sigma^2I)X\{(X^TX)^{-1}\}^T$$

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

employing the rule $(A^{-1})^T = (A^T)^{-1}$

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

Properties of the Least-Squares Estimators

- Using

$$(X^T X)^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}$$

and using $\sum_{i=1}^n x_i^2 = S_{xx} + n\overline{x}^2$

$$Var(b) = \sigma^2 (X^T X)^{-1},$$

We can see that:

$$Var(b_1) = \frac{\sigma^2}{S_{xx}}$$

and

$$Var(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}} \right)$$

Properties of our Estimators

$$E\left(\hat{\beta}_0\right) = \beta_0$$

unbiased

$$E\left(\hat{\beta}_1\right) = \beta_1$$

$$\text{Var}\left(\hat{\beta}_0\right) \downarrow 0 \ as \ n \uparrow \infty$$

$$\text{Var}\left(\hat{\beta}_1\right) \downarrow 0 \ as \ n \uparrow \infty$$