STAT 420

$$\hat{\mathbf{Y}} = \mathbb{X}\hat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} = \mathbb{H}\mathbf{Y}, \quad \text{where } \mathbb{H} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T.$$

$\mathbb{H}$ is called the **hat-matrix** and is the matrix of the orthogonal projection onto the subspace V spanned by columns of matrix $\mathbb{X}$.

$\mathbb{H}_{ii} = h_i$ are called **leverages**. The value of h_i depends only on $\mathbb{X}$, and not $\mathbf{Y}$. Large values of h_i are due to extreme values in $\mathbb{X}$. A point with high leverage has the potential to (greatly) influence the fit (regression equation).

$\sum h_i$ = the number of parameters in the model $= p$. An average value for h_i is $\dfrac{p}{n}$.

A "rule of thumb" is that leverages of more than $2 \cdot \dfrac{p}{n}$ should be looked at more closely.

<u>Example 1</u>:

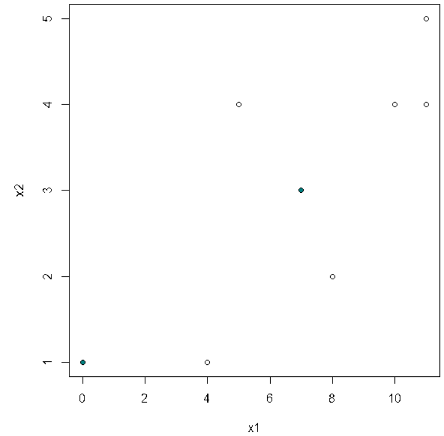
```
> x = c(2,6,8,8,12,16,20,20,22,26)
> y = c(58,105,88,118,117,137,157,169,149,202)
>
> X = cbind(rep(1,10), x)
>
> H = X %*% solve(t(X)%*%X) %*% t(X)
>
> lev = rep(0,10)
> for (i in 1:10) lev[i] = H[i,i]
> lev
 [1] 0.3535211 0.2126761 0.1633803 0.1633803 0.1070423 0.1070423 0.1633803
 [8] 0.1633803 0.2126761 0.3535211
> sum(lev)
[1] 2
```

OR

```
> fit =lm(y ~ x)
> influence(fit)$hat
        1         2         3         4         5         6         7         8 
0.3535211 0.2126761 0.1633803 0.1633803 0.1070423 0.1070423 0.1633803 0.1633803 
        9        10 
0.2126761 0.3535211
```

For Simple Linear Regression $$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}.$$

Example 2:




```
> x1 = c( 0,11,11, 7, 4,10, 5, 8)
> x2 = c( 1, 5, 4, 3, 1, 4, 4, 2)
> y =  c(11,15,13,14, 0,19,16, 8)
>
> plot(x1,x2)
>
> X = cbind( rep(1,8), x1, x2 )
>
> H = X %*% solve(t(X)%*%X) %*% t(X)
> lev = rep(0,8)
> for (i in 1:8) lev[i]=H[i,i]
> lev
[1] 0.6000 0.3750 0.2875 0.1250 0.4000 0.2125 0.5875 0.4125
> sum(lev)
[1] 3
```

OR

```
> fit = lm(y ~ x1 + x2)
> influence(fit)$hat
     1      2      3      4      5      6      7      8
0.6000 0.3750 0.2875 0.1250 0.4000 0.2125 0.5875 0.4125
```

(0, 1)   (7, 3)

```
> lm(y ~ x1 + x2)

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
        3.7         -0.7          4.4


> y[1] = 20          ### point 1 has large leverage
> lm(y ~ x1 + x2)

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1           x2
      8.875       -1.375        4.625
```

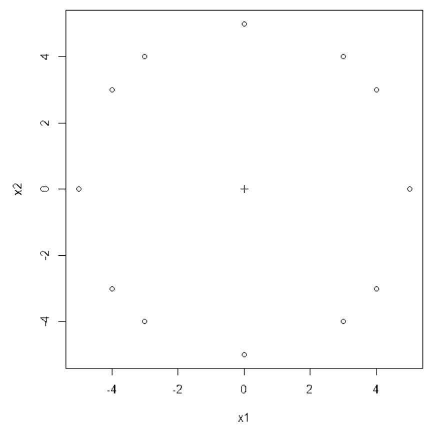
```
> y[1] = 11
> y[4] = 30          ### point 4 has small leverage
> lm(y ~ x1 + x2)

Call:
lm(formula = y ~ x1 + x2)

Coefficients:
(Intercept)           x1            x2
        5.7         -0.7           4.4

> mean(x1); mean(x2)
[1] 7
[1] 3
```

Example 3:




```
> x1 = c(5,4,3,0,-3,-4,-5,-4,-3,0,3,4)
> x2 = c(0,3,4,5,4,3,0,-3,-4,-5,-4,-3)
> par(pty="s")          ### square plot
> plot(x1,x2)
> points(0,0,pch=3)
> X = cbind(rep(1,12),x1,x2)
> H = X %*% solve(t(X)%*%X) %*% t(X)
> lev = rep(0,12)
> for (i in 1:12) lev[i]=H[i,i]
> lev
 [1] 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25 0.25
> sum(lev)
[1] 3
```

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbb{I} - \mathbb{H})\,\mathbf{Y}. \qquad\qquad \mathrm{Var}(e_i) = (1 - h_i)\,\sigma^2.$$

Studentized residuals: $$r_i = \frac{e_i}{s\sqrt{1-h_i}}, \qquad i = 1, 2, \dots, n.$$

Cook's Distance: $$D_i = \frac{1}{p}\, r_i^2 \,\frac{h_i}{1-h_i}, \qquad i = 1, 2, \dots, n.$$

measures the influence of a data point on the regression equation.