# Leverage

- Some cases have high *leverage*, the potential to greatly affect the fit.
- These cases are outliers in the space of predictors.
- Often the residuals for these cases are not large because
  - the response is in line with the other values, or
  - the high leverage has caused the fitted model to be pulled toward the observed response.

# Summary

1. The leverage exerted by the $i$'th case is $h_{ii}$, the $i$'th diagonal element of the hat matrix.
2. Properties:
   - $0 \leq h_{ii} \leq 1$
   - if there is an intercept term in a regression model $h_{ii} \geq \frac{1}{n}$
   - if there are $r$ observations with the same $\boldsymbol{x}$, the leverage for those observations is $\leq 1/r$. (groups of potentially influential cases are masked)
   - a general guideline is to flag cases where $h_{ii} > 2p/n$, where $p$ is the number of columns of $\boldsymbol{X}$, equal to $k+1$ in a multiple regression with $k$ predictors and an intercept.
3. To get leverage values in R, use the command $hat(model.matrix(output))$, where $output$ is the output from a call to $lm$.

- The fitted value at case $i$ is

$$\hat{y}_i = (\boldsymbol{Hy})_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i}^{n} h_{ij} y_j$$

  a linear combination of all the responses.

- Ideally all cases contribute, with those at and closest to $\boldsymbol{x}_i$ dominating.
- In influential cases $h_{ii}$ approaches 1, and $h_{ij}$ approaches 0, for $j \neq i$.
- One can inspect the $h_{ii} = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i$, called *leverage values*, to identify those which are large.
- In R you use the command $hat(model.matrix(output))$ to get the leverage values, where *output* is the output from *lm*.

- Often it is difficult to find a case with high leverage by examining each predictor separately or in pairs using bivariate plots
  - the case may not be extreme in any particular predictor, but still be far from the centroid of the predictors.
  - Recall that $Var(\hat{y}_i) = h_{ii}\sigma^2$ and $Var(e_i) = (1 - h_{ii})\sigma^2$, so cases with high leverage have large estimation variance and small residual variance.
  - In simple linear regression

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

    so the minimum is $1/n$ at $\bar{x}$ and the maximum occurs when $x$ is furthest from $\bar{x}$.
  - More generally, $h_{ii}$ measures the distance of the predictors from their centroid.
  - The sum of the $h_{ii}$ is $tr(\boldsymbol{H}) = k + 1 = p$, so their average is $\bar{h} = (k+1)/n =$ p/n.

# How to identify hidden extrapolation

- Our book points out the danger of hidden extrapolation when predicting (Section 3.8).
- They note that any $\boldsymbol{x}_0$ with $\boldsymbol{x}_0^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 > h_{max}$, where $h_{max}$ is the largest value in the dataset, will imply extrapolation beyond the cases in the dataset.

# $0 \leq h_{ii} \leq 1$

- Because $\boldsymbol{H} = \boldsymbol{HH}$ and $\boldsymbol{H} = \boldsymbol{H}^T$

$$h_{ii} = \sum_{j=1}^{n} h_{ij} h_{ji} = h_{ii}^2 + \sum_{j \neq i}^{n} h_{ij}^2 \tag{1}$$

  so

$$h_{ii}(1 - h_{ii}) \geq 0$$

  and

$$0 \leq h_{ii} \leq 1.$$

- Some statistical packages flag cases where $h_{ii} > 2(k+1)/n = 2p/n$.

# When an intercept $\beta_0$ is included in a multiple regression model $h_{ii} \geq \frac{1}{n}$

- In the notes about adding variables to a regression we partitioned the $\boldsymbol{X}$ matrix into $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and saw that $\boldsymbol{H} = \boldsymbol{H}_1 + \boldsymbol{H}_{2.1}$.
- Let $\boldsymbol{X}_1$ be the vector of 1's, so that $\boldsymbol{H}_1 = \boldsymbol{J}/n$ and

$$\boldsymbol{H}_{2.1} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}^T}$$

  where $\tilde{\boldsymbol{X}} = (\boldsymbol{I} - \boldsymbol{J}/n)\boldsymbol{X}$ contains the deviations of the predictors from their means.
- The $i$th diagonal entry of $\boldsymbol{H}$ is

$$h_{ii} = \frac{1}{n} + [\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T]_{ii}.$$

- The second term is positive so

$$h_{ii} \geq 1/n.$$

# Maximum leverage with replicate $x$

- The second term is of the form

$$\sum_j \sum_k (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\tilde{C}_{jk}$$

  where $\tilde{C}_{jk}$ is the $jk$th entry of $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$, and so measures the distance of the predictors in the $i$th case from the centroid $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_k)^T$ in the $k$ dimensional space of predictors.

- When two cases $i$ and $k$ have the same predictors, $(\boldsymbol{x}_i = \boldsymbol{x}_k)$, (or equivalently, when there are two $y$ values at the same $\boldsymbol{x}$)

$$h_{ik} = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_k = h_{ii}$$

- From equation (1)

$$h_{ii} = 2h_{ii}^2 + \sum_{j \neq i,k}^{n} h_{ij}^2$$

so

$$h_{ii}(1 - 2h_{ii}) \geq 0$$

and

$$0 \leq h_{ii} \leq 1/2.$$

- So, the maximum leverage value is halved when there are two cases with the same values for the predictors.

- More generally, if $r$ cases have the same predictors, (or equivalently, when there are $r$ replicate values of $y$ at $\boldsymbol{x}$), the maximum possible leverage value for these cases is $1/r$.

- Groups of potentially influential cases are *masked*, and cannot be detected by examining the $h_{ii}$.

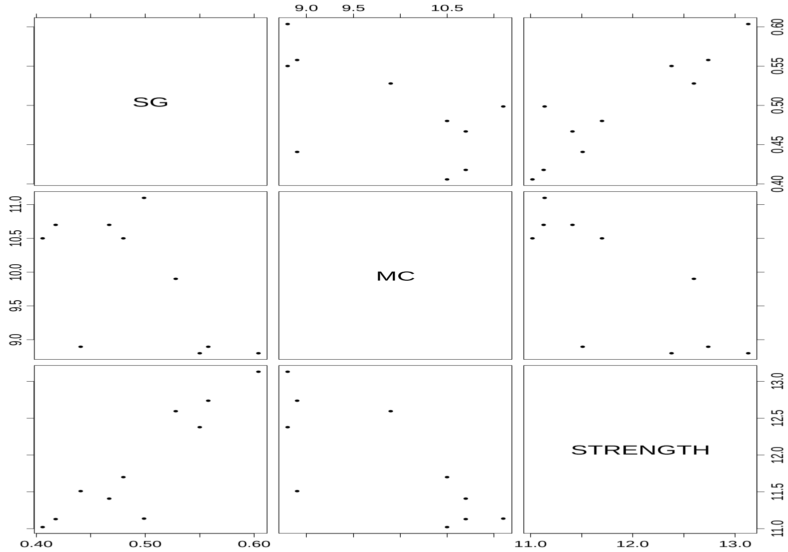
# Example: strength of wood beams

Example: Data on the strength of wood beams was given by Hoaglin and Welsch (*The American Statistician, 1978, vol 32, pp 17-22*). The response is *Strength* and the predictors are *Specific Gravity* and *Moisture Content*. The data are

| Beam Number | Specific Gravity | Moisture Content | Strength |
|---|---|---|---|
| 1 | .499 | 11.1 | 11.14 |
| 2 | .558 | 8.9 | 12.74 |
| 3 | .604 | 8.8 | 13.13 |
| 4 | .441 | 8.9 | 11.51 |
| 5 | .550 | 8.8 | 12.38 |
| 6 | .528 | 9.9 | 12.60 |
| 7 | .418 | 10.7 | 11.13 |
| 8 | .480 | 10.5 | 11.70 |
| 9 | .406 | 10.5 | 11.02 |
| 10 | .467 | 10.7 | 11.41 |

- The correlation matrix of the data is

| | SG | MC | STRENGTH |
|---|---|---|---|
| SG | 1.0000000 | -0.6077351 | 0.9131352 |
| MC | -0.6077351 | 1.0000001 | -0.7592328 |
| STRENGTH | 0.9131352 | -0.7592328 | 1.0000000 |


SG
MC
STRENGTH
0.40
0.50
0.60
9.0
9.5
10.5
11.0
12.0
13.0
0.40
0.45
0.50
0.55
0.60
9.0
9.5
10.0
10.5
11.0
11.5
12.0
12.5
13.0

- There is a positive association between *Strength* and *SG* and a negative association with *MC*.
- There is also a negative association between *Strength* and *MC*, with one value (in the lower left corner) quite different from the others.
- The linear model

$$Strength = \beta_0 + \beta_1 SG + \beta_2 MC + \epsilon$$

gives output as follows.

```
> summary(woodlm.out)

Call: lm(formula = wood.Str ~ wood.SG + wood.MC)
Residuals:
     Min       1Q  Median      3Q    Max
 -0.4442 -0.1278 0.05365 0.1052 0.4499

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)  10.3015   1.8965     5.4319   0.0010
    wood.SG   8.4947   1.7850     4.7589   0.0021
    wood.MC  -0.2663   0.1237    -2.1522   0.0684

Residual standard error: 0.2754 on 7 degrees of freedom
Multiple R-Squared: 0.9
F-statistic: 31.5 on 2 and 7 degrees of freedom, the p-valu
```

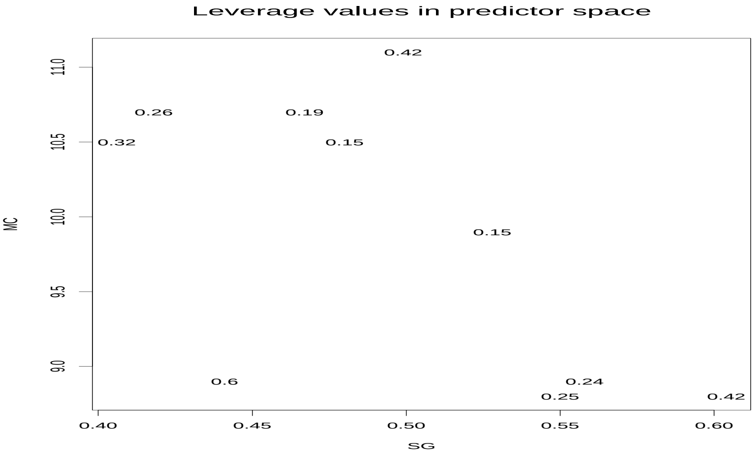
- The leverage values for the linear model in the two predictors are.

```
> hat(model.matrix(woodlm.out))
 [1] 0.4178935 0.2418666 0.4172806 0.6043904 0.2521824 0.14
 [7] 0.2616385 0.1540321 0.3155106 0.1873364
```

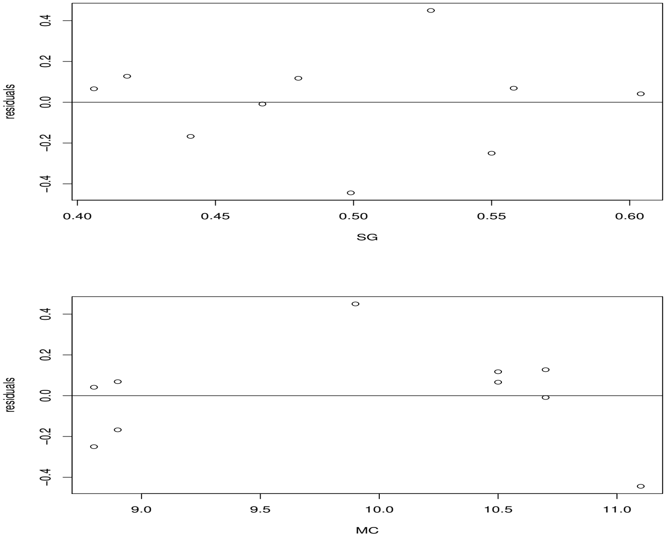
- Case 4 has the largest leverage, this is the case with low $SG$ and $MC$.

- Note that the leverage values sum to 3, and that $2\bar{h} = 2(3)/10 = .6$ so that case 4 would be flagged as high leverage by some packages.

- The leverage values are plotted in the *SG*, *MC* space below.
- One can see how the values increase as you move toward the extremes of the data.



- The residuals from this model are shown below.



- This is only a small data set, but one possible extension to the model is to add a quadratic term in MC.

```
>wood.MC2 = wood.MC^2
>woodlm2.out=lm(wood.Str~wood.SG + wood.MC + wood.MC2)
>summary(woodlm2.out)

Call:
lm(formula = wood.Str ~ wood.SG + wood.MC + wood.MC2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.59511    8.49356  -5.015 0.002416 **
wood.SG       9.68175    0.72851  13.290 1.12e-05 ***
wood.MC      10.42822    1.71125   6.094 0.000889 ***
wood.MC2     -0.54221    0.08672  -6.252 0.000776 ***
---

Residual standard error: 0.1085 on 6 degrees of freedom
Multiple R-Squared: 0.9867,
Adjusted R-squared:  0.98
F-statistic: 148.3 on 3 and 6 DF,  p-value: 5.13e-06
```

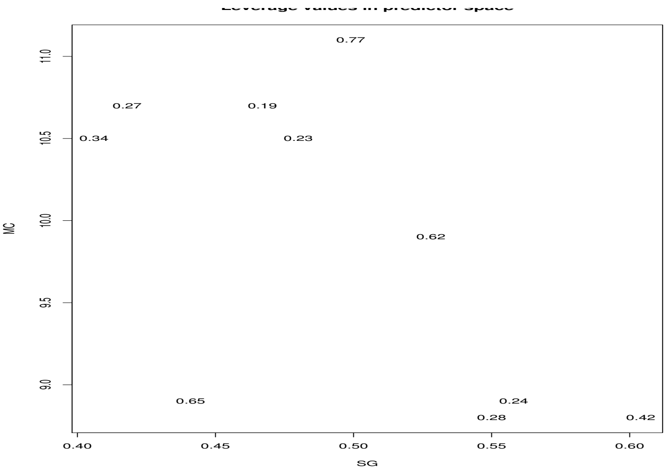
- The fit has been improved ($s$ has been reduced from .2754 to .1085, $R^2$ has increased from .9 to .99) and the $MC^2$ term is highly significant.

- The leverage values change with the model - $\boldsymbol{X}$ has one more column.

```
> hat(model.matrix(woodlm2.out))
 [1] 0.7657191 0.2418690 0.4241376 0.6469168
0.2836093 0.6163116 0.2662545
 [8] 0.2304277 0.3371019 0.1876526
```

- The first case now has the largest leverage value.

- With an extra predictor, however, $2\bar{h} = .8$, so none of these values meet the threshold.


Leverage values in predictor space
MC
SG
11.0
10.5
10.0
9.5
9.0
0.40
0.45
0.50
0.55
0.60
0.77
0.27
0.19
0.34
0.23
0.62
0.65
0.24
0.28
0.42