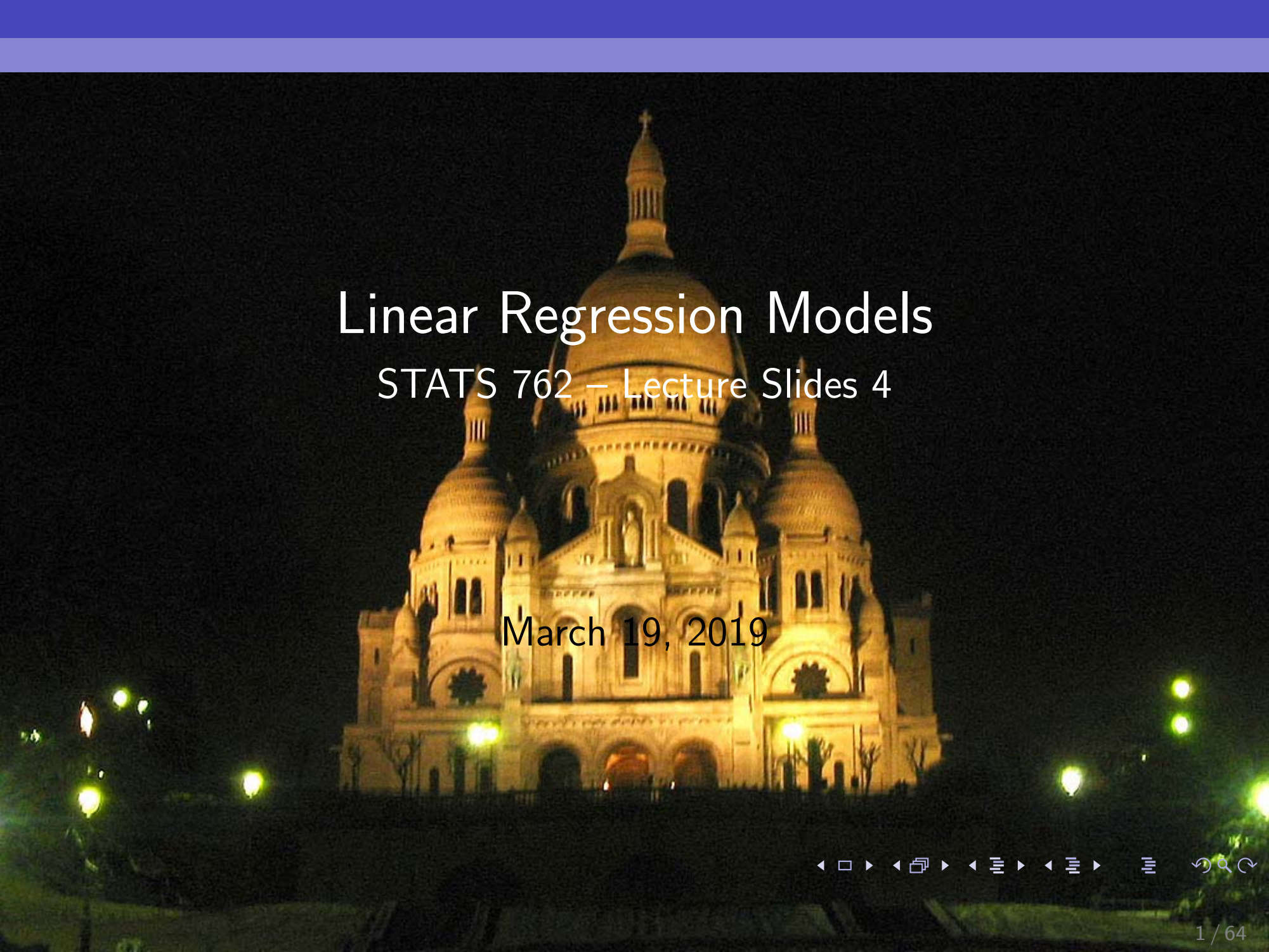


Linear Regression Models

STATS 762 – Lecture Slides 4

March 19, 2019

Diagnostics

We will now shift our attention to diagnostics. Diagnostics are used to identify problems with the data and problems with the fitted model.

- Data problems include outliers, high leverage points (HLP's) and influential points, and multicollinearity.
- Problems with the fitted model occur when the model assumptions (linearity, Normality, independent errors constant variance) are not adequately satisfied.

Data Problems

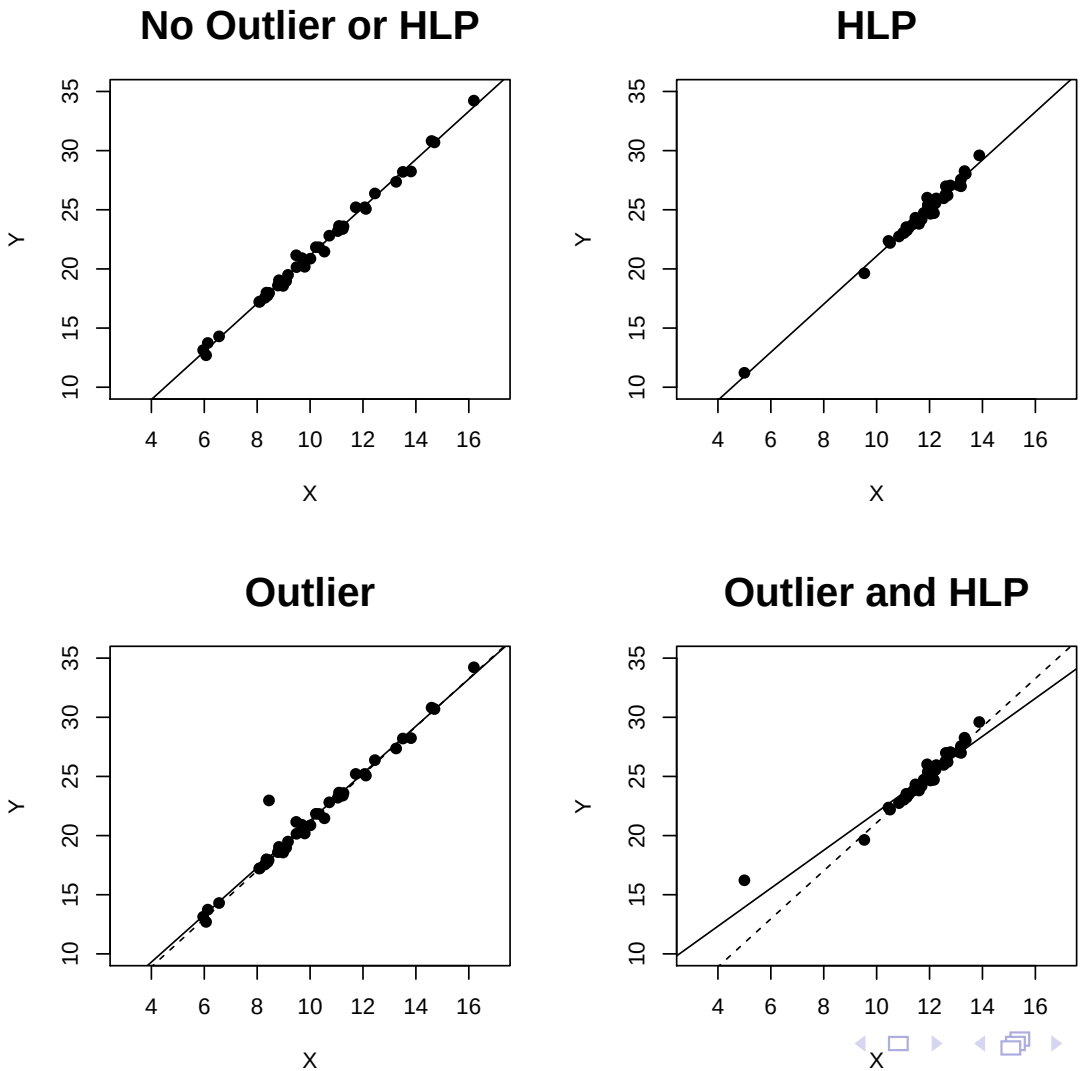
Outliers: observations with large error terms – the value of the response is unusual given the values of the explanatory variables.

High leverage points: observations that have the potential to have a big impact on the fitted model - these points have unusual combinations of values for the regressors.

Influential points: observations that actually do have a big impact on the fitted model – these points are often the result of an observation that is both an outlier and a high leverage point.

Multicollinearity: occurs when there is one or more near linear dependencies among the regressors.

Outliers and High Leverage Points



Detecting HLP's

Our main diagnostic tool for detecting high leverage points are the hat matrix diagonals (HMD's).

$$\underbrace{\begin{bmatrix} \widehat{\mu}_1 \\ \widehat{\mu}_2 \\ \widehat{\mu}_3 \\ \widehat{\mu}_4 \\ \vdots \\ \widehat{\mu}_n \end{bmatrix}}_{\hat{\boldsymbol{\mu}}_{\mathbf{Y}}} = \underbrace{\begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14} & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & h_{24} & \cdots & h_{2n} \\ h_{31} & h_{32} & h_{33} & h_{34} & \cdots & h_{3n} \\ h_{41} & h_{42} & h_{43} & h_{44} & \cdots & h_{4n} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ h_{n1} & h_{n2} & h_{n3} & h_{n4} & \cdots & h_{nn} \end{bmatrix}}_{\mathbf{H}} \underbrace{\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{bmatrix}}_{\mathbf{Y}}$$

Hat Matrix Diagonals

The ith diagonal element of $\mathbf{H}$, h_{ii}, measures the influence or leverage of the ith observation.

$$\widehat{\mu}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n$$

- Thus h_{ii} represents the extent to which y_i determines $\widehat{\mu}_i$.

Properties of the HMD's

The following properties of the HMD's can be deduced:

- $0 \leq h_{ii} \leq 1$
- $\sum_i h_{ii} = k + 1$

Large Hat Matrix Diagonals

How large must h_{ii} be to indicate "excessive leverage"?
From the results on the previous slide we have:

$$\text{average } h_{ii} = (k+1)/n$$

- h_{ii}'s greater than $5(k+1)/n$ or even $3(k+1)/n$ are considered large.
- Or just look at a plot of h_{ii}'s and see if any are unusually large.

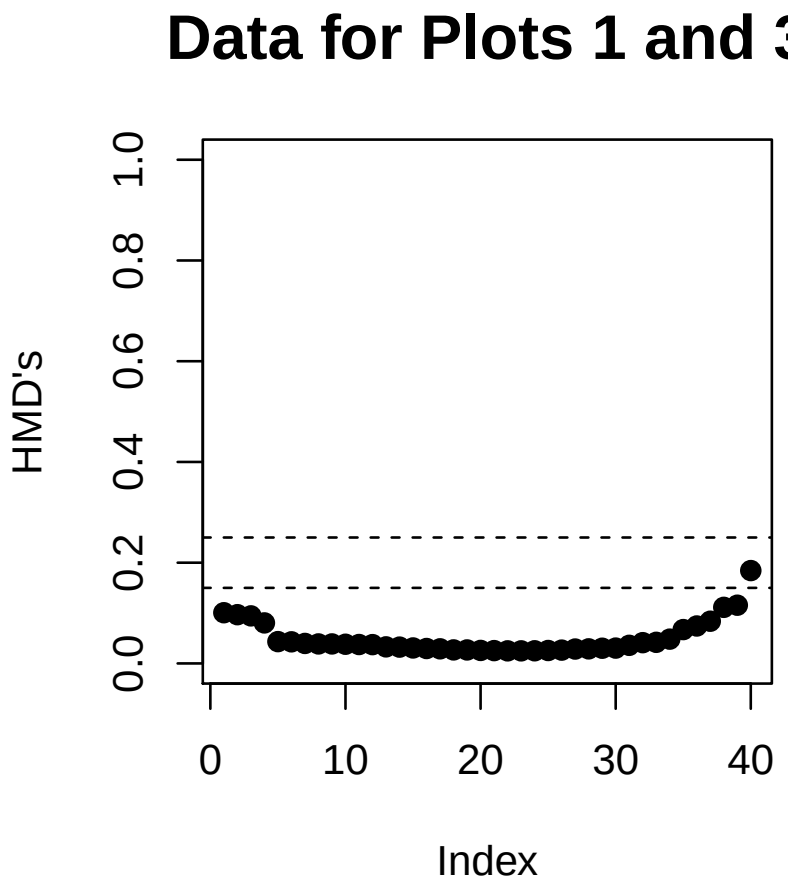
Getting the HMD's

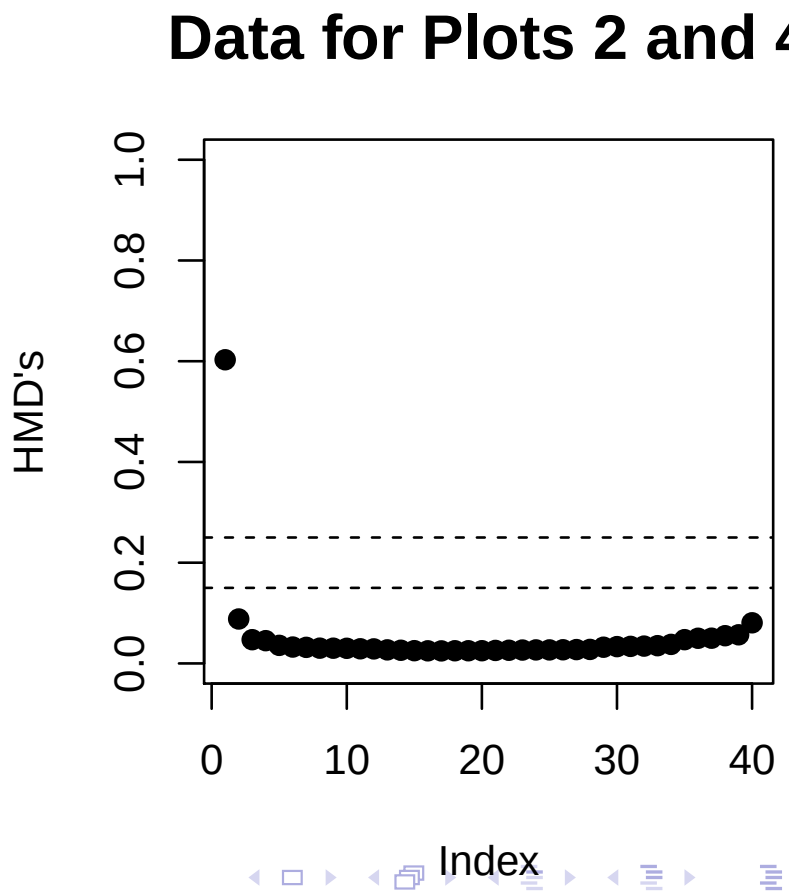
The hat matrix diagonals can be extracted from a `lm` object using `influence`:

```
> hmds<-influence(catheter.lm)$hat
> round(hmds,2)
   1    2    3    4    5    6    7    8    9   10   11   12
0.11 0.51 0.11 0.15 0.12 0.61 0.13 0.39 0.10 0.34 0.12 0.31
```

Index Plots of HMD's

Index plots of the HMD's for our simulated data sets from slide 4.





Improved Residuals for Detecting Outliers

Residuals do not have equal variances:

$$\mathbf{r} \sim \mathrm{N}(\mathbf{0},\ (\mathbf{I} - \mathbf{H})\sigma^2) \quad \text{thus} \quad \mathrm{Var}(r_i) = (1 - h_{ii})\sigma^2.$$

- This "flaw" can be corrected by using standardized residuals:

$$r_i^* = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

  which results in $\mathrm{Var}(r_i') \approx 1$ for all i.
- A further refinement results in the studentized residuals:

$$r_i^\dagger = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$$

  where $\hat{\sigma}_{(i)}$ ignores the i observation in estimating σ.

Getting Residuals

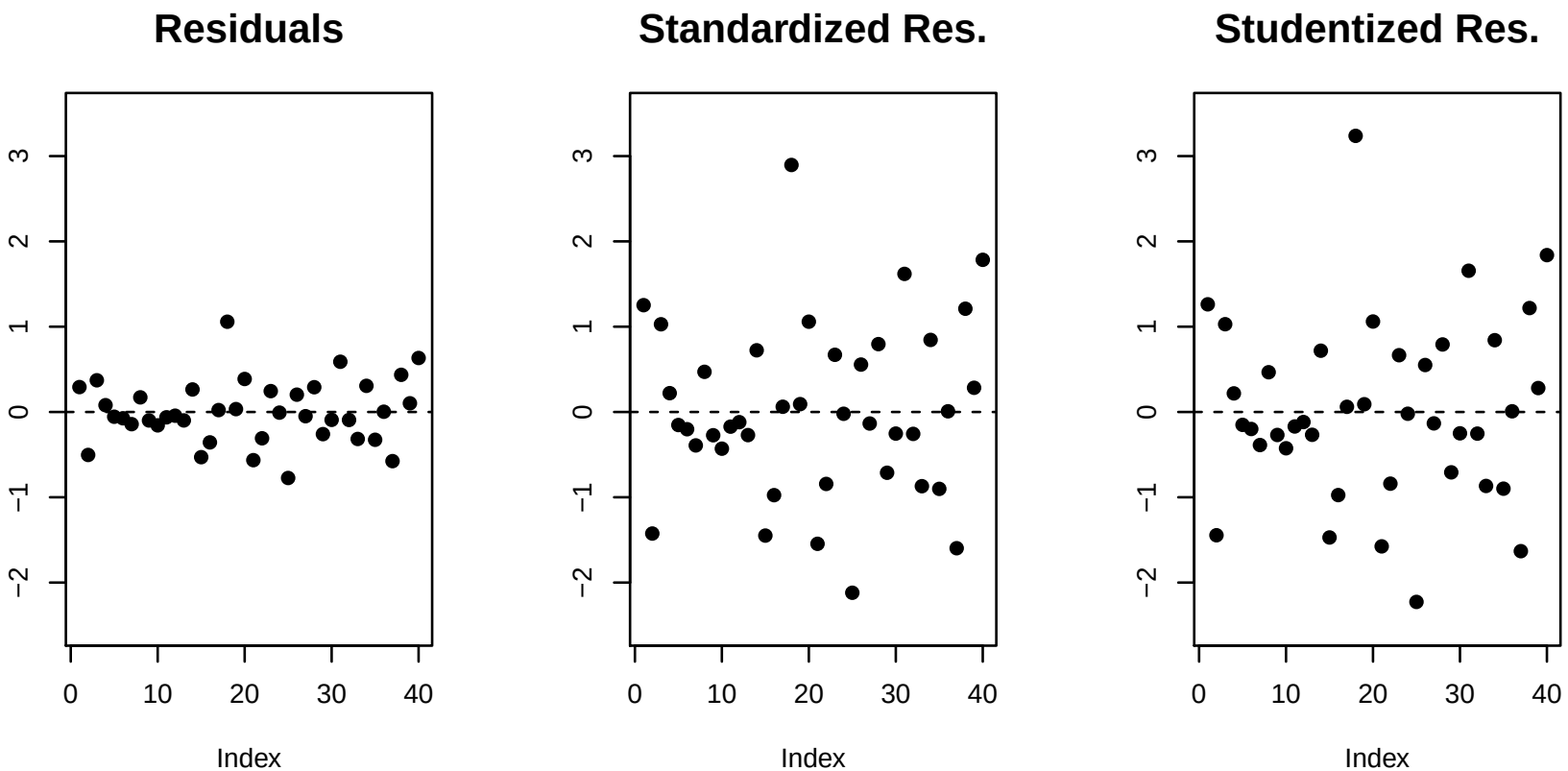
The MASS package for R has functions to extract standardized and studentized residuals from an lm object.

```
> residuals(catheter.lm)
          1           2           3           4
-0.03954422 -1.62559020 -1.06265657  1.56687083  ...
> library(MASS)
> stdres(catheter.lm)
          1           2           3           4
-0.01106713 -0.61254689 -0.29787832  0.45086713  ...
> studres(catheter.lm)
          1           2           3           4
-0.01043427 -0.58994315 -0.28223712  0.42996511  ...
```

In order these are the residuals, the standardized residuals and the studentized residuals (using the terminology in R).

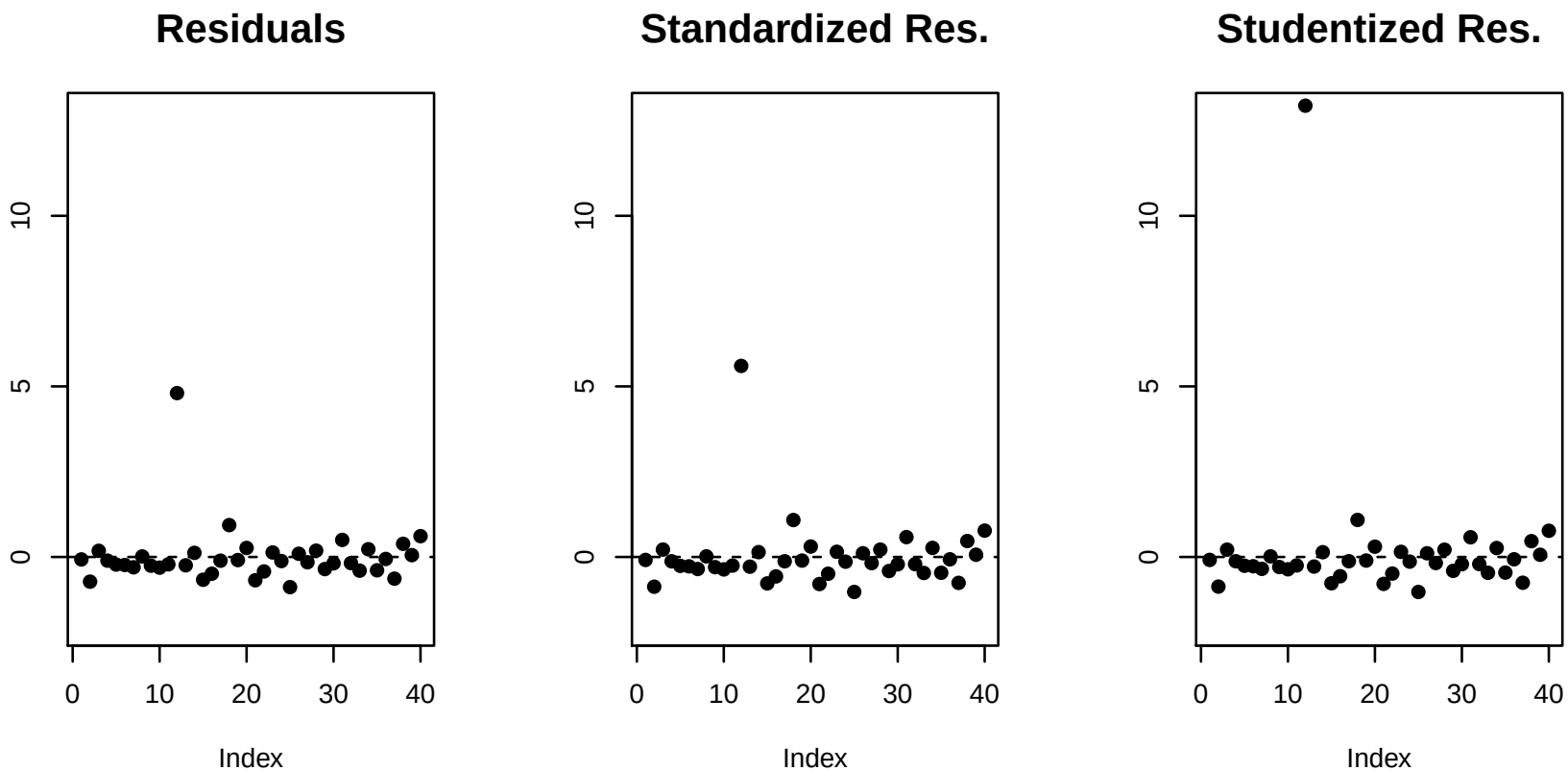
The Three Types of Residuals

The three types of residuals for the "HLP" data (slide 4).



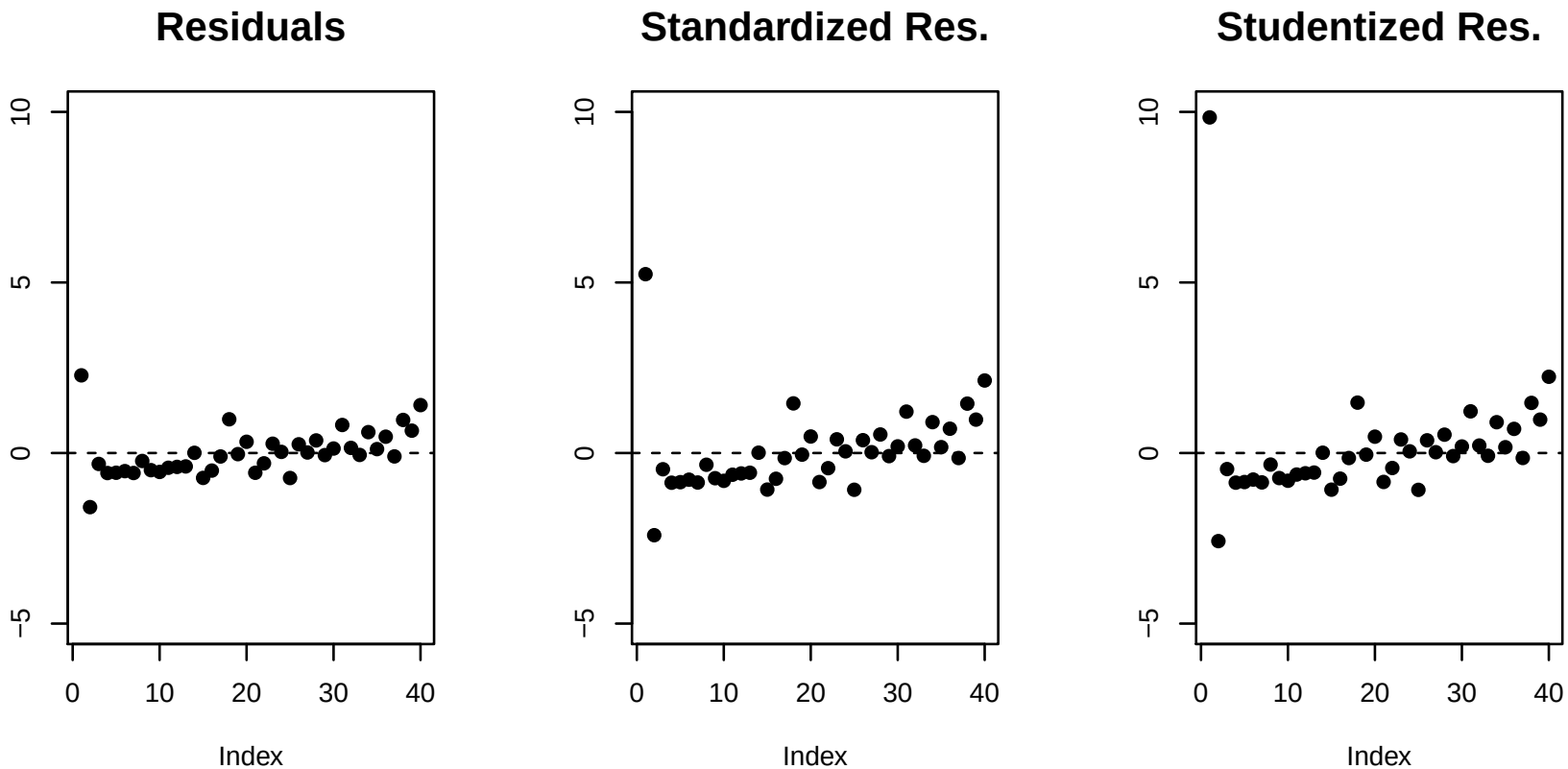
The Three Types of Residuals (cont.)

The three types of residuals for the "Outlier" data (slide 4).



The Three Types of Residuals (cont.)

The three types of residuals for the "Outlier and HLP" data (slide 4).



Detecting Influential Points

Observations that are both high leverage points and outliers will be influential.

- A plot of the HMD's is useful for identifying high leverage points.
- A plot of (ordinary) residuals may miss an outlier if it has high leverage. In such cases, a plot of standardized residuals or a plot of studentized residuals is much more effective.

Cook's Distance

Cook's distance is a direct measure of influence. The idea is to consider the over all change in $\widehat{\boldsymbol{\beta}}$ when the i observation is removed.

$$D_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})\mathbf{X}^t\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)})}{(k+1)\hat{\sigma}^2}$$

where $\widehat{\boldsymbol{\beta}}_{(-i)}$ is the fitted vector of parameters when the observation i is deleted.

Cook's distance can also be expressed as

$$D_i = \frac{r_i^{*2}}{k+1} \times \frac{h_{ii}}{1-h_{ii}}$$

Cook's Distance

Cook's distance is an overall measure of the changes in the $\hat{\beta}_j$'s when the ith observation is deleted.

- It combines information from r_i^* and h_{ii}.
- Plot the D_i's and see if any are unusually large.
- It has been suggested that any D_i's bigger than $F_{k+1,n-k-1}(.50) \approx 1$ are "large."

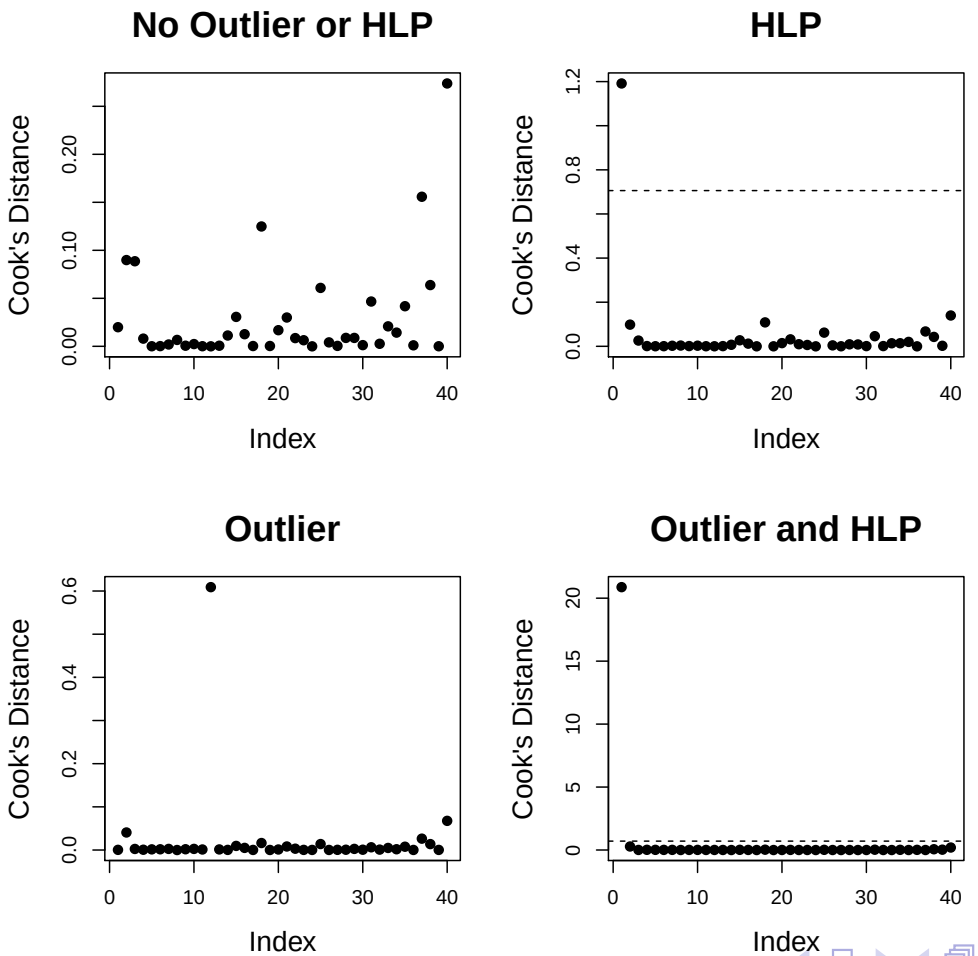
Getting Cook's Distance

To extract Cook's distances using *R*:

```
> cooks.distance(catheter.lm)
           1            2            3            4
4.823040e-06 1.284435e-01 3.599878e-03 1.233285e-02  ...
```

Cook's Distance Plots

Cook's distance plots for our simulated data sets from slide 4.



Diagnostic Plots

We can get diagnostic plots for an `lm` object using:

`plot(my.lm,which=1:6).`

These plots are:

1. Residuals vs fitted values.
2. A Normal Q-Q plot of the standardized residuals.
3. Square root of the absolute values of the standardized residuals vs fitted values.
4. An index barchart of Cook's distance.
5. Standardized residuals vs leverage (HMD's).
6. Cook's distance vs leverage/(1-leverage).

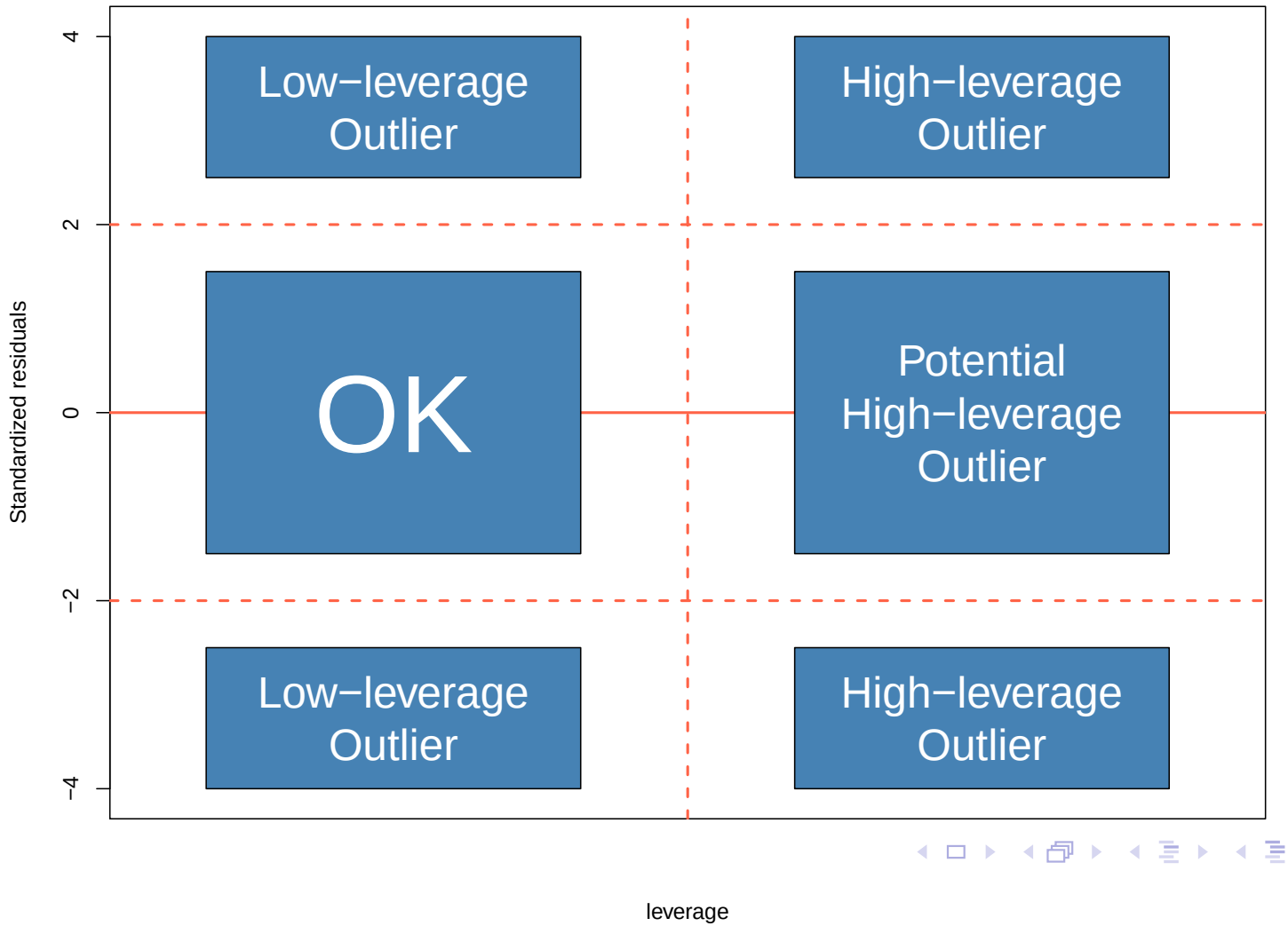
If you just use `plot(my.lm)` plots 1, 2, 3 and 5 are produced.

Standardized Residuals vs Leverage Plot

The standardized residuals vs leverage is particularly useful for identifying high leverage and influential observations.

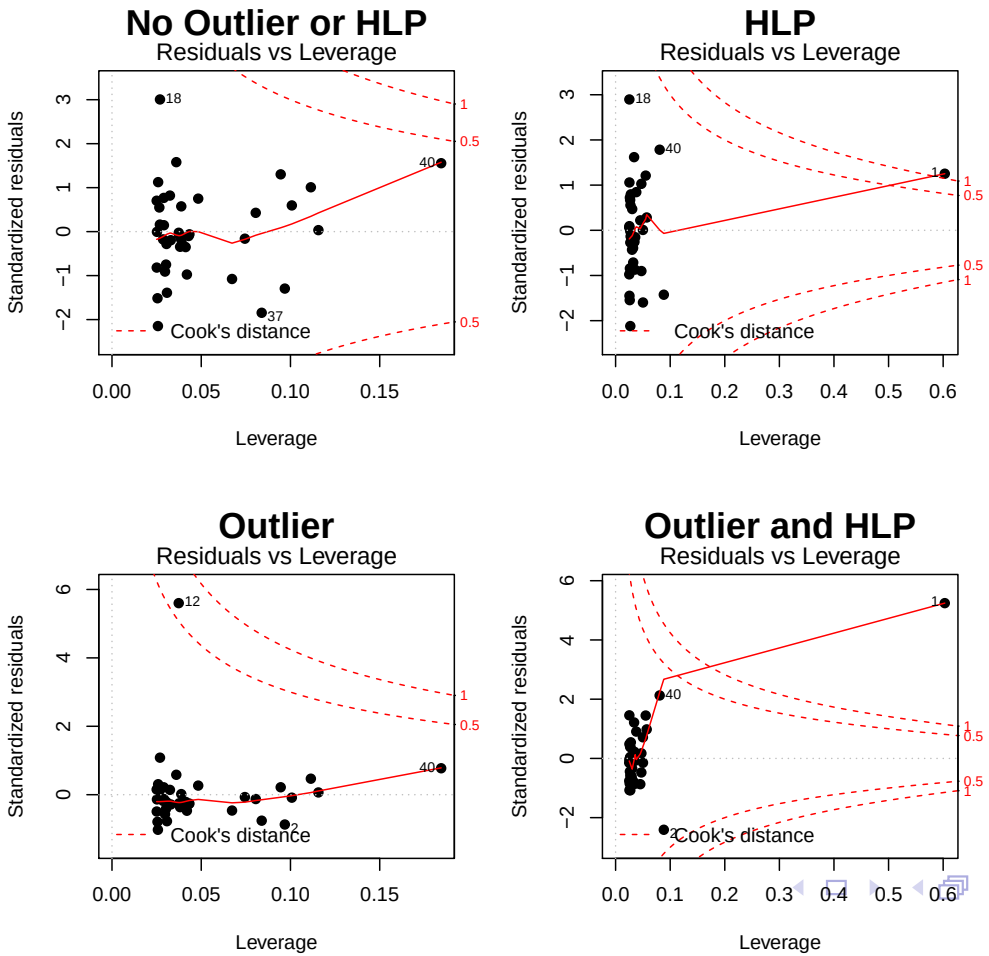
- Outliers show up as extreme points at the top or bottom of the plot.
- Observations with high leverage show up as extreme points near the right edge of the plot.
- Influential points show up near the upper right or lower right corners of the plot. Contour lines corresponding to Cook's distances of 1 and of 0.5 are given to help determine influential points.

Interpreting LR plots



Standardized Residuals vs Leverage Plots

Standardized residuals vs leverage plots for our simulated data sets from slide 4.



Example: Education Expenditure Data

Data for 50 states of the USA

Variables are:

- educ: Per capita expenditure on education (response).
- percap: Per capita income.
- under18: Number of residents per 1000 under 18.
- urban: Number of residents per 1000 in urban areas.

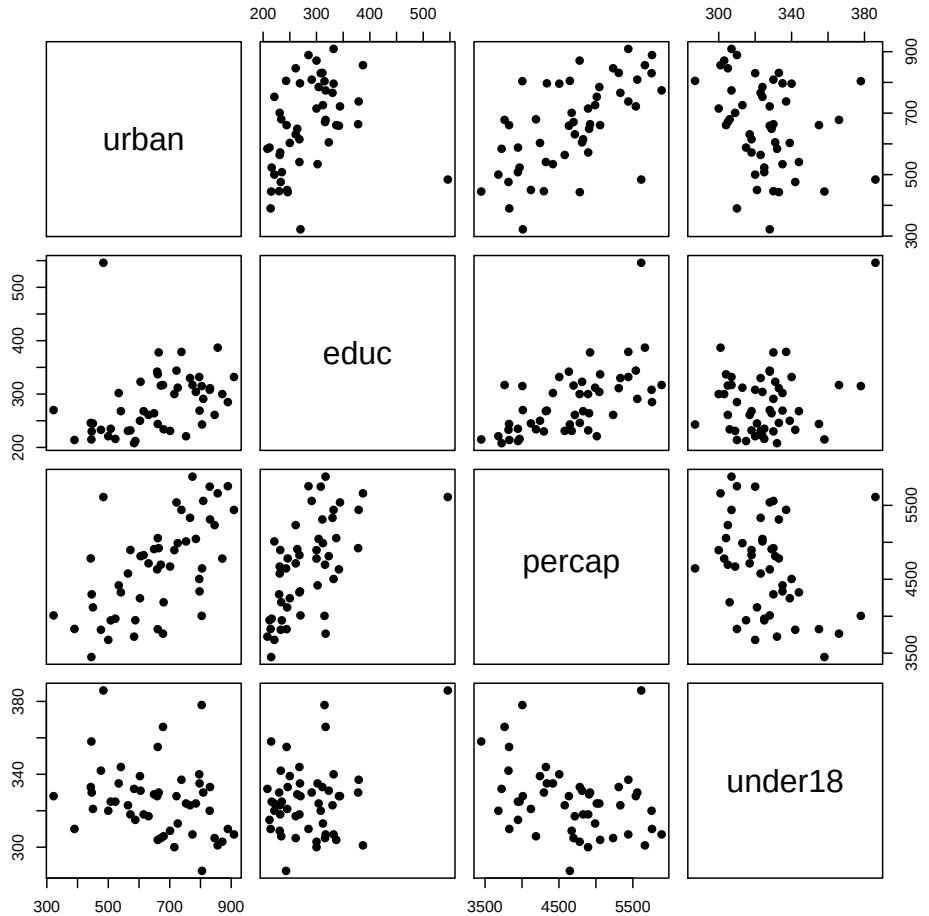
Fit model:

```
> lm(educ ~ percap + under18 + urban,data=educ.df)
```

Data Exploration

```
> head(educ.df)
  urban educ percap under18
1   508  235   3944      325
2   564  231   4578      323
3   322  270   4011      328
4   846  261   5233      305
5   871  300   4780      303
6   774  317   5889      307
> summary(educ.df)
     urban            educ          percap        under18     
 Min.   :322.0   Min.   :208.0   Min.   :3448   Min.   :287.0  
 1st Qu.:546.8   1st Qu.:234.2   1st Qu.:4137   1st Qu.:310.8  
 Median :662.5   Median :269.5   Median :4706   Median :324.5  
 Mean   :657.4   Mean   :284.6   Mean   :4676   Mean   :325.7  
 3rd Qu.:782.2   3rd Qu.:316.8   3rd Qu.:5054   3rd Qu.:333.0  
 Max.   :909.0   Max.   :546.0   Max.   :5889   Max.   :386.0  
```

Pairs Plot


urban
educ
percap
under18
200
300
400
500
300
340
380
900
700
500
300
200
300
400
500
5500
4500
3500
380
340
300
300
500
700
900
3500
4500
5500

One Unusual Point

One observation is showing up as having an unusually large per capita expenditure on education.

```
> educ.df[educ.df$educ>450,]
   urban educ percap under18
50   484  546   5613     386
```

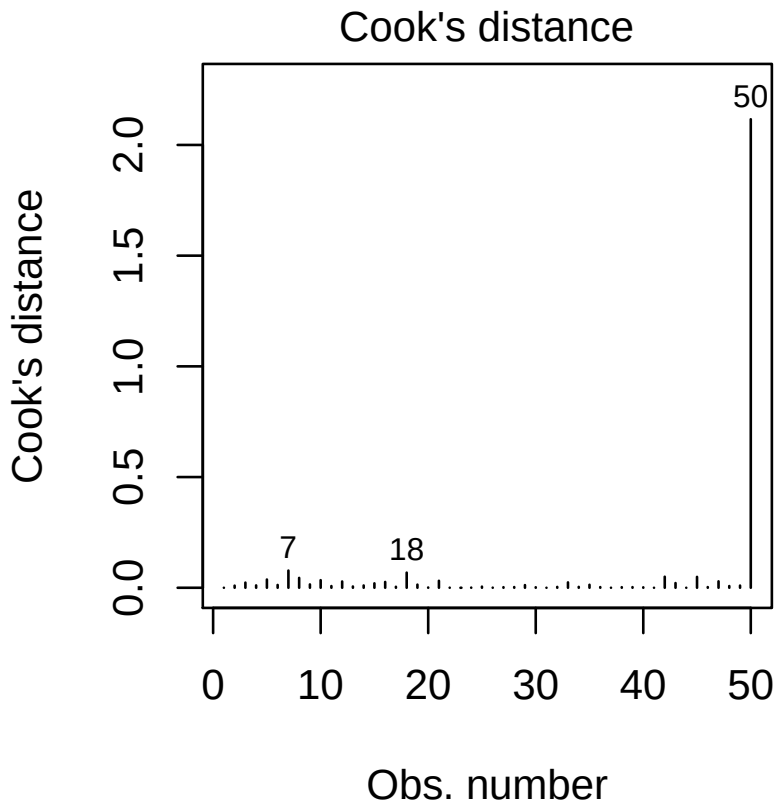
It also has a high values of `under18` and `percap` and a low value of `urban`.

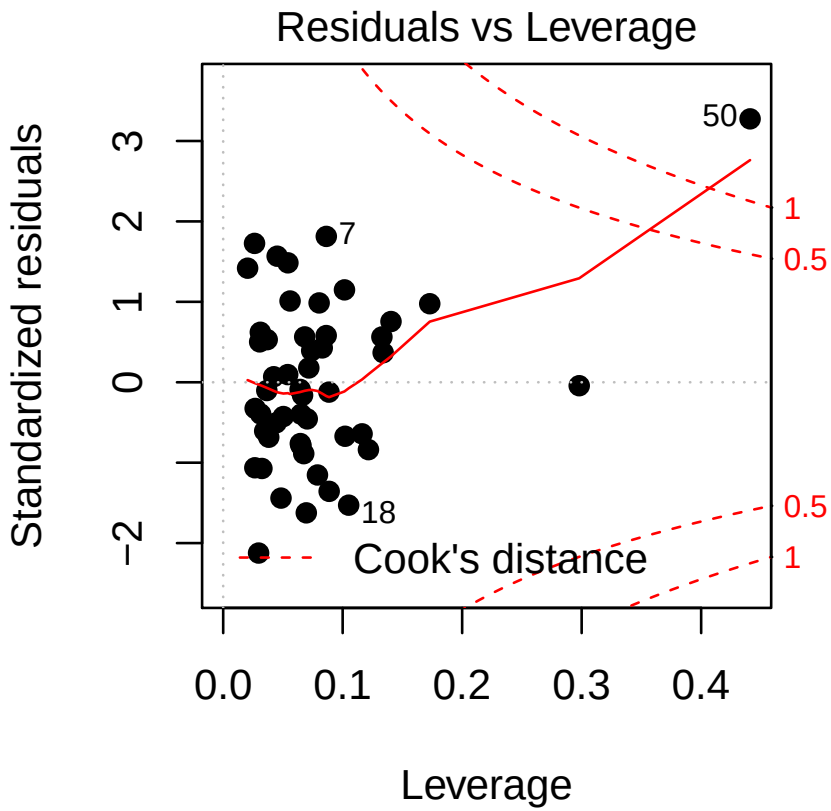
Which state is it?

Initial Fit

```
> educ.lm<-lm(educ ~ percap + under18 + urban,data=educ.df)
> summary(educ.lm)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -555.92562  123.46634  -4.503 4.56e-05 ***
percap         0.07236    0.01165   6.211 1.40e-07 ***
under18        1.55134    0.31545   4.918 1.16e-05 ***
urban         -0.00476    0.05174  -0.092    0.927
---
Residual standard error: 40.53 on 46 degrees of freedom
Multiple R-squared:  0.5902,Adjusted R-squared:  0.5634
F-statistic: 22.08 on 3 and 46 DF,  p-value: 5.271e-09
```

Diagnostic Plots





What do we learn?

Assessing the Impact

Observation 50 is clearly influential. The most direct way of assessing its impact is to delete it and see how the fitted model changes.

```
> educ2.lm<-lm(educ ~ percap + under18 + urban,
                                        data=educ.df[-50,])
> summary(educ2.lm)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -278.06430  132.61422  -2.097 0.041664 *
percap         0.04827    0.01220   3.958 0.000266 ***
under18        0.88983    0.33159   2.684 0.010157 *
urban          0.06624    0.04966   1.334 0.188948
---
Residual standard error: 35.88 on 45 degrees of freedom
Multiple R-squared:  0.4947,Adjusted R-squared:  0.461
F-statistic: 14.68 on 3 and 45 DF,  p-value: 8.365e-07
```

Changes

Deleting this one point has significantly altered the fitted model.

- The coefficients for both `percap` and `under18` have both decreased in magnitude.
- The coefficients for `urban` has increased in magnitude and switched sign.
- `Residual standard error` has decreased.
- `Multiple R-squared` and `Adjusted R-squared` have both decreased.

The last point seems a bit odd – explanations?

Diagnosing Data Problems for GLM's

For GLM's the procedures for diagnosing data related problems are similar to those for ordinary regression.

- There are more types of residuals.
- The hat matrix diagonals are defined with respect to the final iteration of the IRLS procedure.
- The definition of Cook's distance is modified as a result of the previous point.

Residuals for GLM's

The `residuals` function can be used to extract 5 different types of residuals from a `glm` object.

Deviance residuals measure the contribution of each observation to the residual deviance for the model.

Pearson residuals are the difference between the response and the fitted value divided by the standard error (similar to the standardised residuals for ordinary regression).

Working residuals are the difference between the constructed variables and the estimated linear predictor for the last step of the IRLS procedure.

Response residuals are the difference between the response and the fitted value.

Partial residuals are working residuals when one of the regressors has been removed (get k sets of residuals in a matrix).

Residuals for GLM's (cont.)

The deviance residuals and the Pearson residuals are the most useful for identifying problems with the data.

- These residuals take the place of the standardized residuals in the diagnostic procedures that were described for ordinary regression.
- Both deviance and Pearson residuals can be standardized by dividing by $\sqrt{1 - h_{ii}}$.
- For logistic regression applied to ungrouped data, residuals (of any sort) have little diagnostic value.

Deviance Residuals

The residual deviance can be written as

$$\text{deviance} = \sum d_i^2$$

where d_i^2 represents the contribution of observation i.

- d_i^2 is the difference in $2 \log f(y_i)$ between the maximal model and the specified model.
- The deviance residual for observation i is $\pm\sqrt{d_i}$ where the residual is positive if $y_i > \widehat{\mu}_i$ and negative otherwise.
- Unusually large values of $\|d_i\|$ indicate observations that the model fits poorly.

Deviance Residuals for Logistic Regression

For grouped data

$$d_i^2 = 2r_i \log\left(\frac{r_i}{n_i\widehat{\pi}_i}\right) + 2(n_i - r_i)\log\left(\frac{n_i - r_i}{n_i - n_i\widehat{\pi}_i}\right)$$

where for the i covariate pattern there are r_i successes out of n_i trials and $\widehat{\pi}_i$ is the estimated probability of success under the logistic model. Define

$$d_i = \pm\left[2r_i \log\left(\frac{r_i}{n_i\widehat{\pi}_i}\right) + 2(n_i - r_i)\log\left(\frac{n_i - r_i}{n_i - n_i\widehat{\pi}_i}\right)\right]^{1/2}$$

where the sign is positive if $r_i > n_i\widehat{\pi}_i$ and negative otherwise.

Deviance Residuals for Logistic Regression

For ungrouped data, the previous expression can be adjusted by setting $n_i = 1$ and $r_i = 0$ or 1. Thus we get

$$d_i = -\sqrt{-2\log(1-\widehat{\pi}_i)} \quad \text{if } y_i = 0$$

$$d_i = \sqrt{-2\log\widehat{\pi}_i} \quad \text{if } y_i = 1$$

- Thus for $y_i = 0$ or 1, the d_i's are completely determined by $\widehat{\pi}_i$'s.

Deviance Residuals for Poisson Regression

For Poisson regression

$$d_i^2 = 2\left[y_i \log\left(y_i/\widehat{\mu}_i\right) - \left(y_i - \widehat{\mu}_i\right)\right]$$

where $\widehat{\mu}_i$ is the estimate of μ_i under the Poisson regression model.

Thus we get

$$d_i = \pm\sqrt{2\left[y_i \log\left(y_i/\widehat{\mu}_i\right) - \left(y_i - \widehat{\mu}_i\right)\right]}$$

where the sign is positive if $y_i > \widehat{\mu}_i$ and negative otherwise.

Pearson Residuals

Pearson residuals are defined as

$$P_i = \frac{y_i - \widehat{\mu}_i}{\widehat{\sigma}_i}$$

where $\widehat{\sigma}_i$ is the estimated standard deviation for observation i.

- $\widehat{\sigma}_i$ usually depends on $\widehat{\mu}_i$.

Pearson Residuals for Logistic Regression

For grouped data, the Pearson residual for the ith covariate pattern is

$$P_i = \frac{r_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

For ungrouped data, this formula becomes

$$P_i = \frac{1-\hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \qquad \text{when } y_i = 1$$

$$P_i = \frac{-\hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \qquad \text{when } y_i = 0$$

Pearson Residuals for Poisson Regression

For Poisson regression the Pearson residuals are given by

$$P_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}.$$

Hat Matrix Diagonals for GLM's

For GLM's the hat matrix is taken from the final iteration of the IRLS procedure:

$$H = \mathbf{W}^{1/2}\mathbf{X}\left(\mathbf{X}^t\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{W}^{1/2}$$

- The hat matrix diagonals (HMD's) in this case are interpreted in the same way as for ordinary regression – measure the leverage of an observation.
- The weights in $\mathbf{W}$ depend on the $\widehat{\mu}_i$'s and the derivative of the inverse link function as well as the values of the regressors.
- h_{ii}'s greater than $5(k+1)/n$ or even $3(k+1)/n$ are considered large.

Cook's Distance for GLM's

For GLM's Cook's distance is defined as:

$$D_i = \frac{P_i^2}{k+1} \times \frac{h_{ii}}{(1-h_{ii})^2}.$$

This is the second equation from slide 17 with the standardized Pearson residuals taking the place of the standardized residuals.

Deviance Change

Another useful leave-one-out diagnostic is the Deviance Change.

- If the ith covariate pattern is left out, the change in the deviance is approximately

$$(\text{Dev. Res})^2 + (\text{Pearson Res})^2 HMD/(1 - HMD)$$

- Identify usually large values using an index plot.
- Big if more than about 4

Diagnostic Plots

Similar diagnostic plots are produced for a `glm` object using:

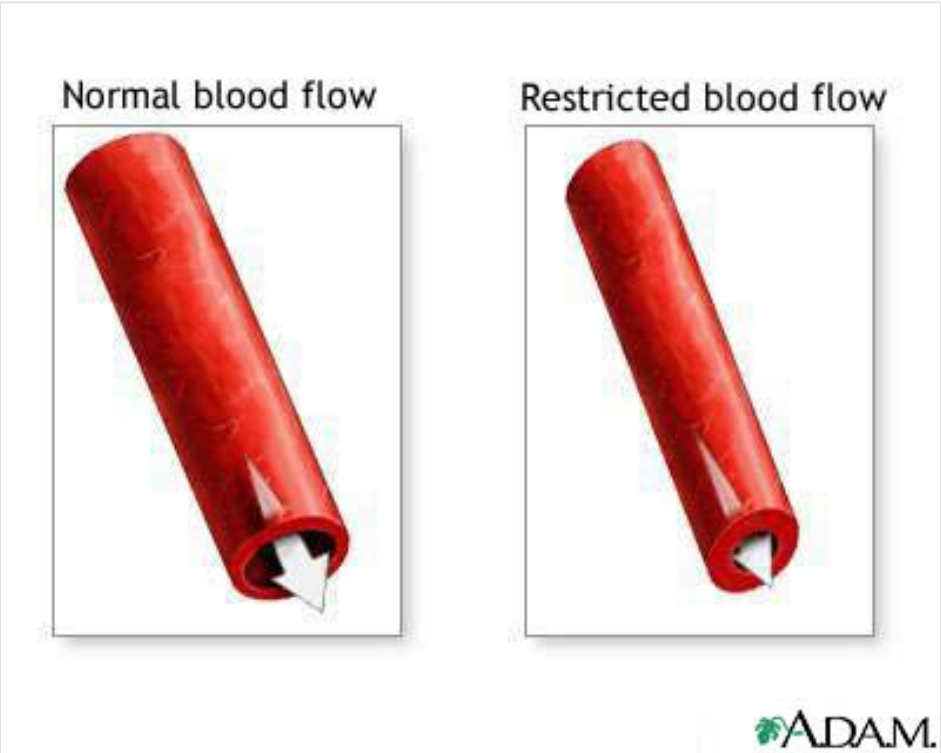
```
plot(my.glm,which=1:6)
```

as were produced for an `lm` object.

1. Deviance residuals vs linear predictors.
2. A Normal Q-Q plot of the standardized deviance residuals.
3. Square root of the absolute values of the standardized deviance residuals vs linear predictors.
4. An index barchart of Cook's distance.
5. Standardized Pearson residuals vs leverage (HMD's).
6. Cook's distance vs leverage/(1-leverage).

Example: Vaso-Constriction Data

Data from a study of reflex vaso-constriction (narrowing of the blood vessels) of the skin of the fingers, which can be caused by a sharp intake of breath



Example: vaso-constriction data (cont)

Variables measured:

Response: 1 = vaso-constriction occurs
0 = doesn't occur

Volume: volume of air breathed in

Rate: rate of intake of breath

Data

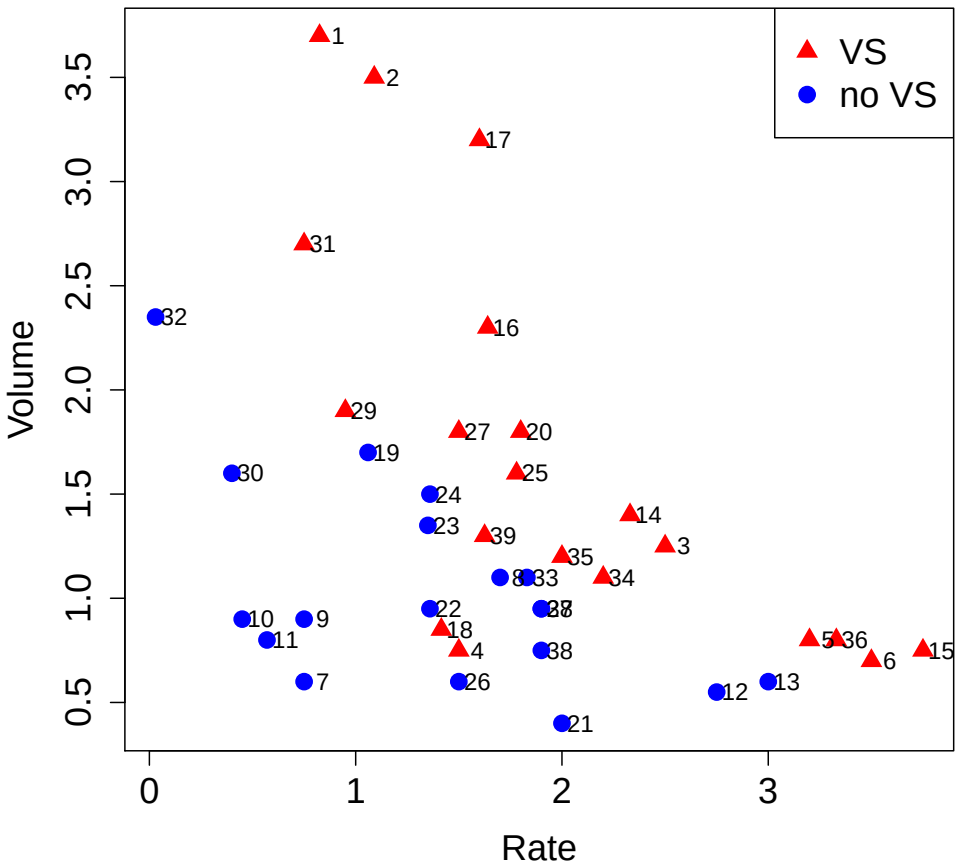
	Volume	Rate	Response
1	3.70	0.825	1
2	3.50	1.090	1
3	1.25	2.500	1
4	0.75	1.500	1
5	0.80	3.200	1
6	0.70	3.500	1
7	0.60	0.750	0
8	1.10	1.700	0
9	0.90	0.750	0
10	0.90	0.450	0
11	0.80	0.570	0
12	0.55	2.750	0
13	0.60	3.000	0
. . .	39 obs in all		

Plotting the Data

To get a plot with red triangles representing vaso-constriction and blue circles representing no vaso-constriction.

```
plot(Volume~Rate,data=vaso.df,
    cex=1.5, cex.axis=1.5, cex.lab=1.5,
    col=ifelse(vaso.df$Response==1, "red","blue"),
    pch=ifelse(vaso.df$Response==1,17,19),
                          xlab="Rate", ylab="Volume")
legend("topright", col=c("red","blue"),pch=c(17,19),
                        legend = c("VS","no VS"),cex=1.5)
text(vaso.df$Rate+.09,vaso.df$Volume,1:39)
```

Plot of Vaso-Constriction Data



Logistic Regression Model

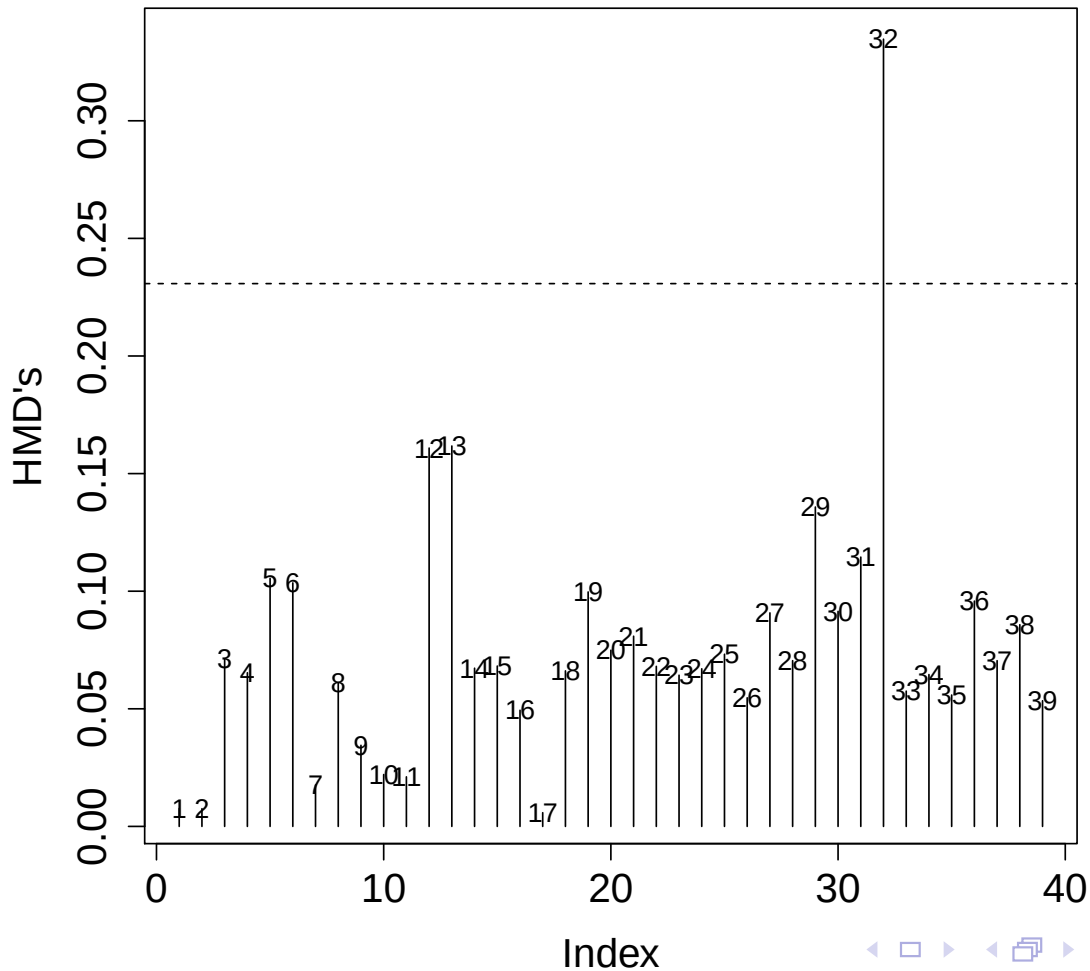
```
> vaso.glm = glm(Response~Rate+Volume,
                 family=binomial, data=vaso.df)
> summary(vaso.glm)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.50657  -0.73464   0.03997   0.48854   2.32935

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5296     3.2332  -2.947  0.00320 **
Rate          2.6491     0.9142   2.898  0.00376 **
Volume        3.8822     1.4286   2.717  0.00658 **
---
    Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 29.772  on 36  degrees of freedom
```

Plotting the HMD's

The HMD's can be extracted form a `glm` object using the `hatvalues` function.

```
vaso.glm = glm(Response~log(Rate)+log(Volume),
  family=binomial, data=vaso.df)
HMD<-hatvalues(vaso.glm)
plot(HMD,ylab="HMD's",type="h", cex=1.5,
  cex.axis=1.5, cex.lab=1.5)
text(HMD)
abline(h=3*3/39, lty=2)
```

Plotting the HMD's (cont.)

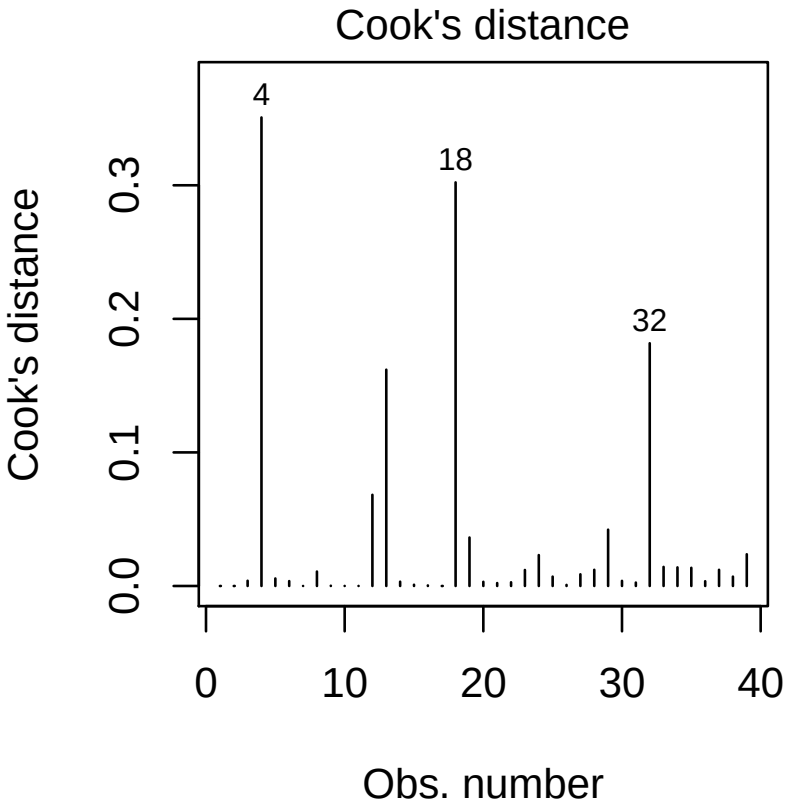


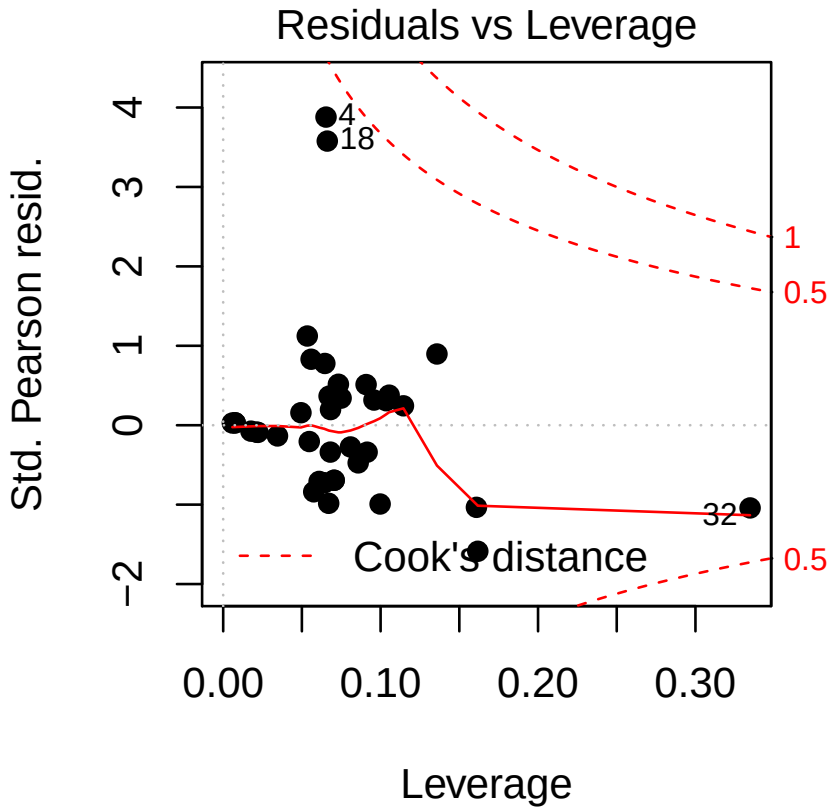
Hat Matrix Diagonals

In ordinary regression, the hat matrix diagonals measure the leverage ("outlying-ness") of the covariates pattern for an observation. The plot on slide 51 would seem to indicate that points 1, 2 and 17 should have large HMD values but this is not the case.

- In logistic regression, the HMD's measure leverage but are down-weighted according to the estimated probability for the observation. The weights are small if the probability is close to 0 or 1.
- Points 1, 2 and 17 had very small weights since the estimated probabilities are close to 1 for these points.
- Point 32 is in a region where the probability is closer to .5 and thus has a bigger weight.

Cook's Distance





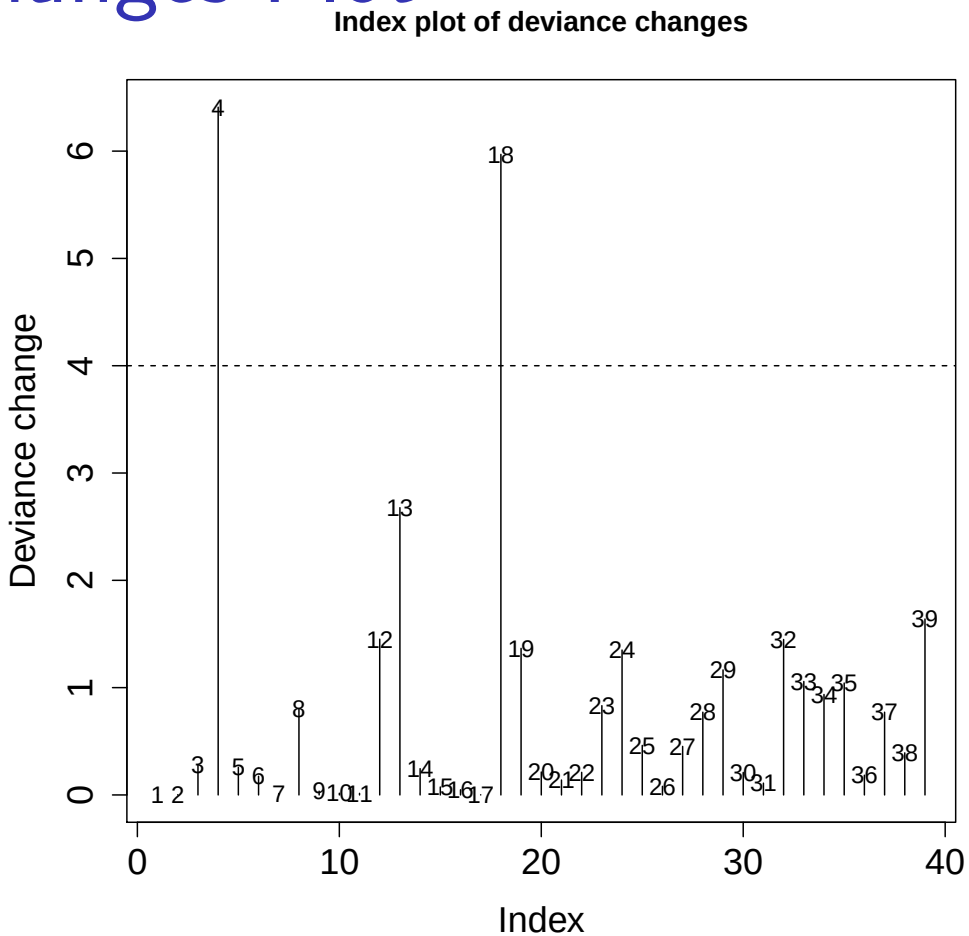
- Observations 4 and 18 are showing up as having the most influence.

Creating a Deviance Changes Plot

Deviance changes are estimated using the formula on slide 45.

```
HMD<-hatvalues(vaso.glm)
dev.r<-residuals(vaso.glm,type="deviance")
pear.r<-residuals(vaso.glm,type="pearson")
Dev.change<-dev.r^2 + pear.r^2*HMD/(1-HMD)
plot(Dev.change,ylab="Deviance change",
type="h",cex=1.5,cex.axis=1.5, cex.lab=1.5,
main="Index plot of deviance changes")
text(Dev.change)
bigdev=4
abline(h=bigdev, lty=2)
```

Deviance Changes Plot



Delete Points?

How much impact are points 4, 18 and 32 having?

- We can delete each in turn and examine changes in the coefficients

Deleting:	None	4	18	32	All 3
(Intercept)	-9.530	-14.338	-13.796	-9.349	-41.989
Rate	2.649	3.790	3.680	2.616	10.744
Volume	3.882	5.897	5.627	3.775	17.495

Big change when all 3 deleted.

- Deleting these points will increase the magnitude of the estimated regression coefficients and may overstate the predictive ability of the model.

Multicollinearity

Multicollinearity is a near linear dependency between the explanatory variables (the X-values are not very well spread out).

- Affects the stability of the fitted regression plane – applies to GLM's as well as ordinary regression.
- Inflates the standard errors in the estimated coefficients
  - may find that a number of coefficients are not significant but if you remove any one of these variables the others become significant

Detecting Multicollinearity

If we have k explanatory variables, then the variance of $\widehat{\beta}_j$ is

$$\text{Var}(\widehat{\beta}_j) = \frac{\sigma^2/(n-1)}{\text{Var}(X_j)(1-R_j^2)}$$

where R_j^2 is the R^2 value when X_j is treated as the response and regressed against the remaining explanatory variables.

- If X_j is orthogonal to all of the remaining explanatory variables then $R_j^2 = 0$ and

$$\text{Var}(\widehat{\beta}_j) = \frac{\sigma^2/(n-1)}{\text{Var}(X_j)}.$$

Variance Inflation Factors (VIF's)

The jth variance inflation factor is defined as

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

It represents the amount that $\text{Var}(\widehat{\beta}_j)$ is inflated due to the correlation between X_j and the remaining regressors.

Calculating VIF's

The VIF's can be calculated fro the correlation matrix for the explanatory variables.
For the education expenditure data (slide 25), the VIF's can be calculated as follows:

```
> Xmat<-model.matrix(educ.lm)[,-1]
> diag(solve(cor(Xmat)))
  percap  under18    urban
1.682571 1.119863 1.677396
```

- values of 1 indicate a regressor that is orthogonal to all other regressors.
- values of 5 or more indicate multicollinearity.
- values of approximately 10 or more indicate serious multicollinearity.

Dealing with Collinearity

It's useful to identify multicollinearity in a data set as it alerts us to potential problems in estimating the impact of certain regressors on the response.

- Multicollinearity indicates an overlap of information (confounding) between two or more of the regressors.
- Can be a serious problem if a goal of the analysis is to evaluate the impact of one of the affected regressors on the response – the data may not be well suited to that purpose.
- Less of a problem if the goal is to find a good predictive model for the response – in this case subset selection methods may well eliminate one or more of the regressors involved.