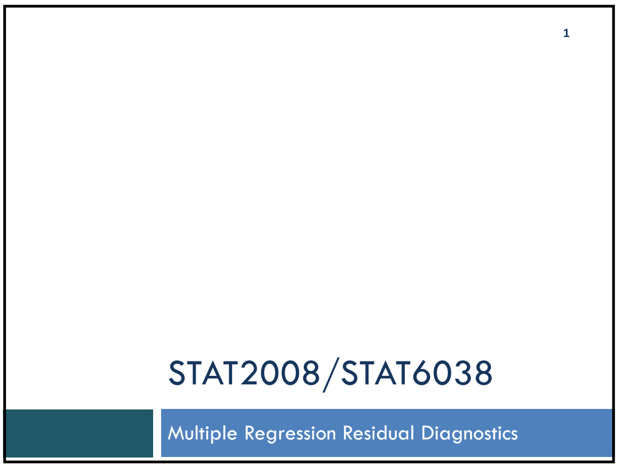
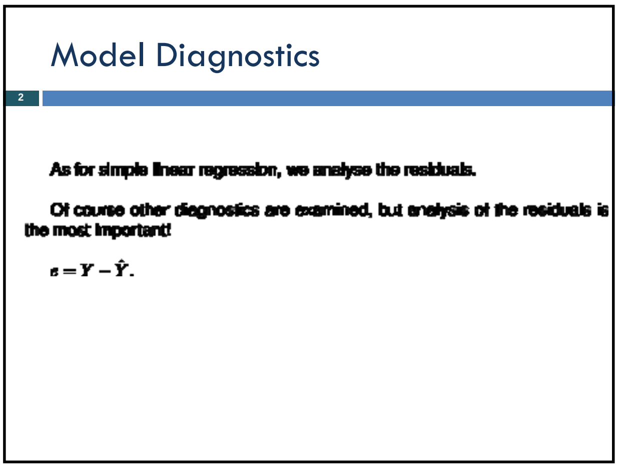


STAT2008/STAT6038

Multiple Regression Residual Diagnostics

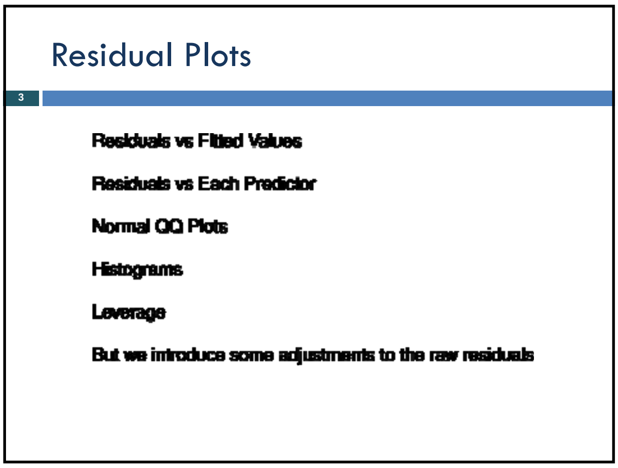


Model Diagnostics

As for simple linear regression, we analyse the residuals.

Of course other diagnostics are examined, but analysis of the residuals is the most important!

$e = Y - \hat{Y}$.



Residual Plots

Residuals vs Fitted Values

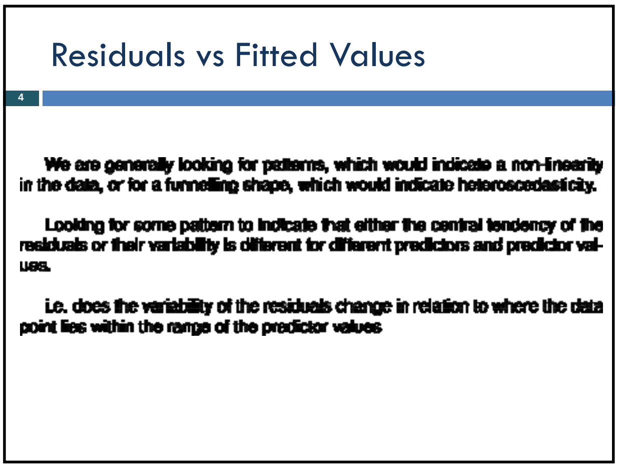
Residuals vs Each Predictor

Normal QQ Plots

Histograms

Leverage

But we introduce some adjustments to the raw residuals

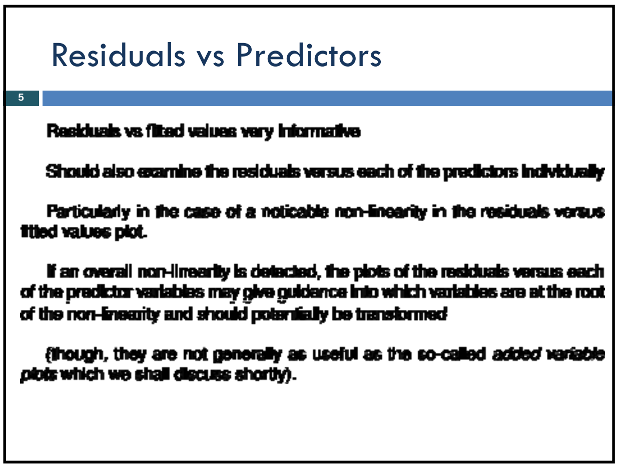


Residuals vs Fitted Values

We are generally looking for patterns, which would indicate a non-linearity in the data, or for a funnelling shape, which would indicate heteroscedasticity.

Looking for some pattern to indicate that either the central tendency of the residuals or their variability is different for different predictors and predictor values.

i.e. does the variability of the residuals change in relation to where the data point lies within the range of the predictor values



Residuals vs Predictors

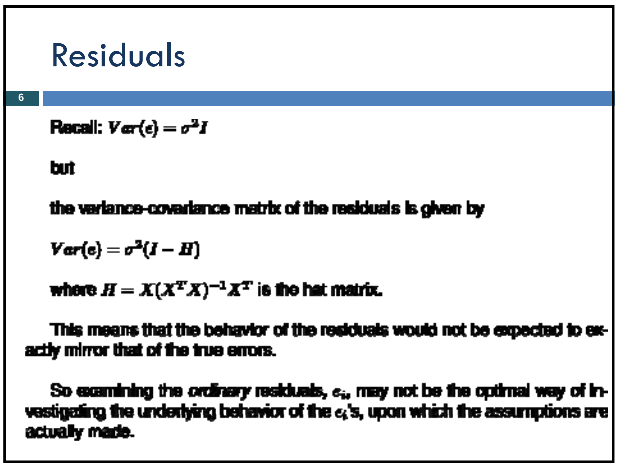
Residuals vs fitted values very informative

Should also examine the residuals versus each of the predictors individually

Particularly in the case of a noticable non-linearity in the residuals versus fitted values plot.

If an overall non-linearity is detected, the plots of the residuals versus each of the predictor variables may give guidance into which variables are at the root of the non-linearity and should potentially be transformed

(though, they are not generally as useful as the so-called *added variable plots* which we shall discuss shortly).



Residuals

Recall: $Var(\epsilon) = \sigma^2 I$

but

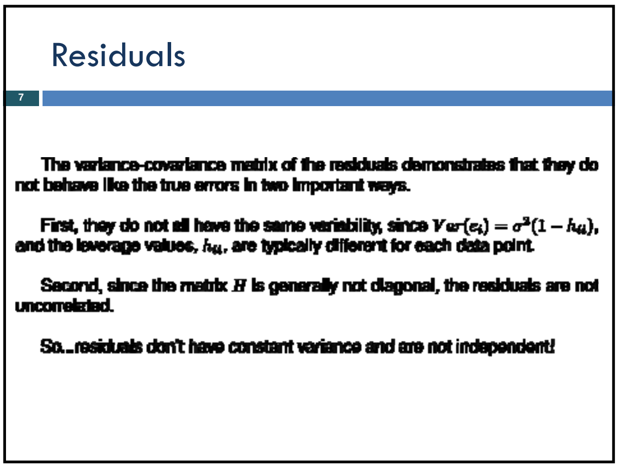
the variance-covariance matrix of the residuals is given by

$Var(e) = \sigma^2(I - H)$

where $H = X(X^T X)^{-1} X^T$ is the hat matrix.

This means that the behavior of the residuals would not be expected to exactly mirror that of the true errors.

So examining the *ordinary* residuals, e_i, may not be the optimal way of investigating the underlying behavior of the ϵ_i's, upon which the assumptions are actually made.



Residuals

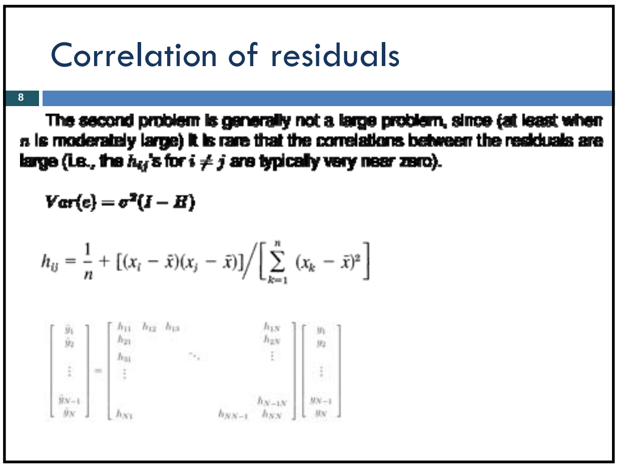
7

The variance-covariance matrix of the residuals demonstrates that they do not behave like the true errors in two important ways.

First, they do not all have the same variability, since $Var(e_i) = \sigma^2(1 - h_{ii})$, and the leverage values, h_{ii}, are typically different for each data point.

Second, since the matrix H is generally not diagonal, the residuals are not uncorrelated.

So...residuals don't have constant variance and are not independent!



Correlation of residuals

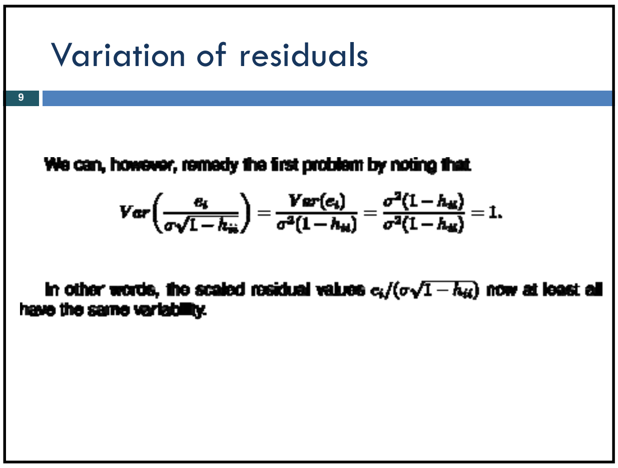
8

The second problem is generally not a large problem, since (at least when n is moderately large) it is rare that the correlations between the residuals are large (i.e., the h_{ij}'s for $i \neq j$ are typically very near zero).

$$Var(e) = \sigma^2(I - H)$$

$$h_{ij} = \frac{1}{n} + [(x_i - \bar{x})(x_j - \bar{x})] \Big/ \left[\sum_{k=1}^{n} (x_k - \bar{x})^2\right]$$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \\ \vdots \\ \\ \hat{y}_{N-1} \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & & & h_{1N} \\ h_{21} & & & & & h_{2N} \\ h_{31} & & & \ddots & & \vdots \\ \vdots & & & & & \\ & & & & & \\ & & & & & h_{N-1N} \\ h_{N1} & & & & h_{NN-1} & h_{NN} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \\ \vdots \\ \\ y_{N-1} \\ y_N \end{bmatrix}$$



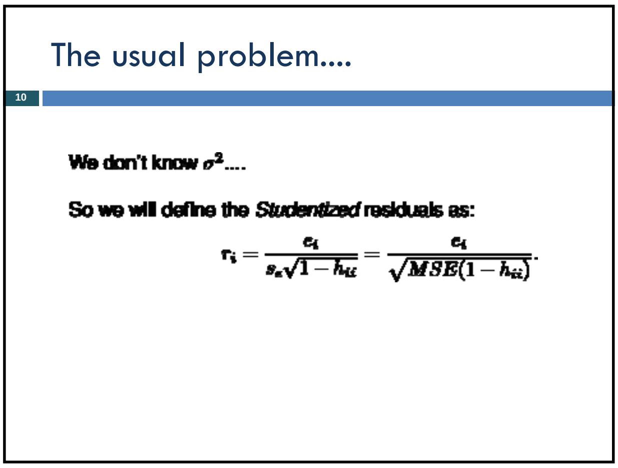
Variation of residuals

9

We can, however, remedy the first problem by noting that

$$Var\left(\frac{e_i}{\sigma\sqrt{1-h_{ii}}}\right) = \frac{Var(e_i)}{\sigma^2(1-h_{ii})} = \frac{\sigma^2(1-h_{ii})}{\sigma^2(1-h_{ii})} = 1.$$

In other words, the scaled residual values $e_i/(\sigma\sqrt{1-h_{ii}})$ now at least all have the same variability.



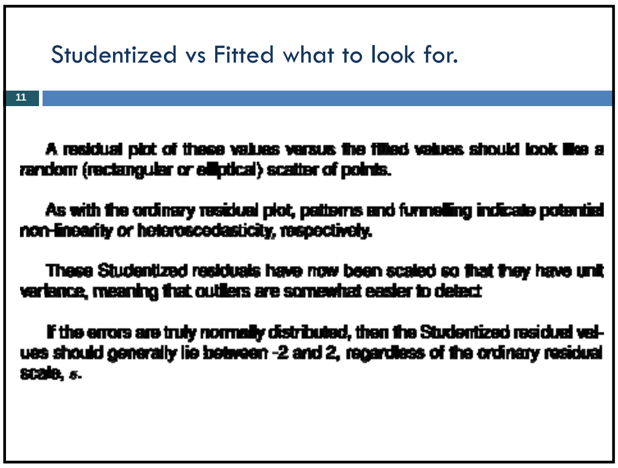
The usual problem....

10

We don't know σ^2....

So we will define the *Studentized* residuals as:

$$r_i = \frac{e_i}{s_e\sqrt{1-h_{ii}}} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}.$$



Studentized vs Fitted what to look for.

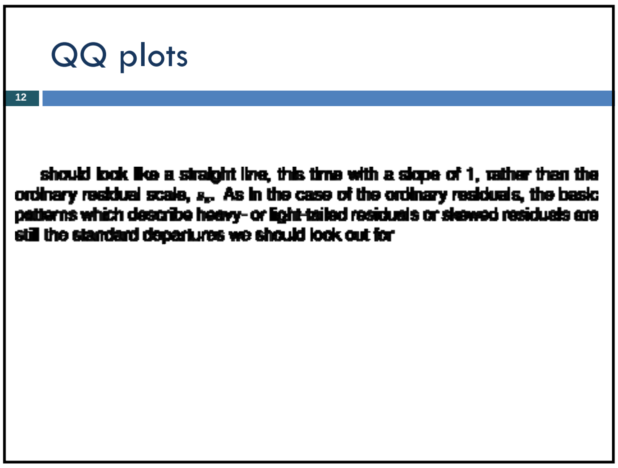
11

A residual plot of these values versus the fitted values should look like a random (rectangular or elliptical) scatter of points.

As with the ordinary residual plot, patterns and funnelling indicate potential non-linearity or heteroscedasticity, respectively.

These Studentized residuals have now been scaled so that they have unit variance, meaning that outliers are somewhat easier to detect

If the errors are truly normally distributed, then the Studentized residual values should generally lie between -2 and 2, regardless of the ordinary residual scale, s.



QQ plots

12

should look like a straight line, this time with a slope of 1, rather than the ordinary residual scale, s_e. As in the case of the ordinary residuals, the basic patterns which describe heavy- or light-tailed residuals or skewed residuals are still the standard departures we should look out for

Outliers

13

The two most common sources of outliers are:

1. There is a location (mean) shift at the i^{th} data point:

$E(Y|X = x_i) = \beta x_i + \Delta$ so that $E(e_i) = \Delta_i \neq 0$.

vs

$E(Y|X = x_i) = \beta x_i$

2. There is a scale shift at the i^{th} data point, so that $Var(e_i) > \sigma^2$.

Outliers.

14

These discrepancies may be caused by something “real”(i.e., related to the underlying nature of the populations under study) or simply by a data collection error.

Be careful! Outliers can contain important information

Perhaps our model was “wrong” in the first place, and the outliers are not really outliers!

Also, we should take care to ensure that our model does not solely represent the central majority of our observed data (i.e., an “overfit” of the data), and instead gives a useful description of the overall population of interest.

**Model Uncertainty – Not assessable but an important concept to be aware of.

PRESS

15

We may argue since every data point has some influence on the resulting fitted regression equation, we should measure the degree to which data points are outliers by using the so-called *PRESS (PREdiction Sum of Squares) residual*:

$$e_{i,-i} = Y_i - \hat{Y}_{i,-i},$$

$\hat{Y}_{i,-i}$ is the predicted value at x_i from a regression with the i^{th} point removed.

$e_{i,-i}$ measures how far the i^{th} response is from a prediction over which it has no influence.

Note that, if there is a location shift, Δ_i, at the i^{th} data point then

$$E(e_{i,-i}) = \Delta_i \neq E(e_i).$$

A useful relationship

16

It might at first seem time-consuming to calculate all the $e_{i,-i}$'s since they apparently each require fitting a new regression.

However, an interesting relation exists between the PRESS residuals and the ordinary residuals:

$$e_{i,-i} = \frac{e_i}{1 - h_{ii}},$$

where h_{ii} is just the usual leverage value for the i^{th} data point in the original regression with all the data points included.

Still need to standardise

17

But we should still standardize the PRESS residuals, since they do not all have the same variance:

$$Var(e_{i,-i}) = Var\left(\frac{e_i}{1 - h_{ii}}\right)$$

$= \frac{1}{(1-h_{ii})^2} Var(e_i) = \frac{\sigma^2(1-h_{ii})}{(1-h_{ii})^2} = \frac{\sigma^2}{1-h_{ii}}$.

So, standardizing the PRESS residuals yields: $\frac{e_{i,-i}}{\sqrt{Var(e_{i,-i})}} = \frac{e_i}{(1-h_{ii})\sqrt{Var(e_{i,-i})}}$ $= \frac{e_i}{(1-h_{ii})\sqrt{\sigma^2/(1-h_{ii})}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$.

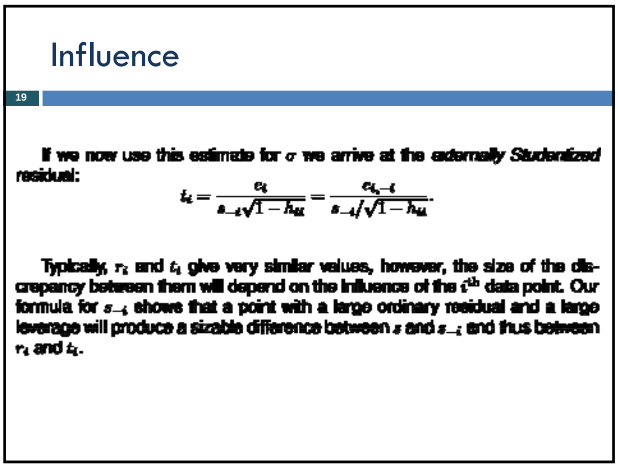
So, once we replace σ by s_e, we see that the standardized PRESS residuals are the same as the Studentized residuals, adding credibility to the use of the r_i's as outlier diagnostics.

Adjusted regression scale

18

In standardizing the PRESS residuals, we used s_e, the regression scale from the original regression on the entire dataset, to estimate σ, and doing so led us to the (*internally*) Studentized residual. However, perhaps a better estimate of σ in this case might be the residual scale from the regression calculated without the i^{th} data point, denoted by s_{-i}, which can be calculated as:

$$s_{-i} = \sqrt{\frac{(n-p)s_e^2 - \{e_i^2/(1 - h_{ii})\}}{n - p - 1}}.$$



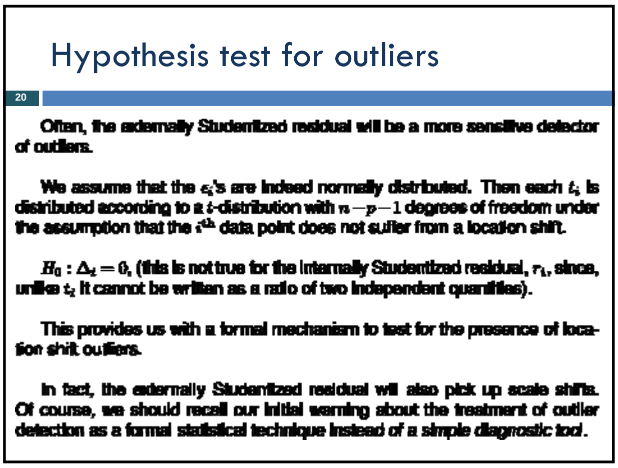
Influence

19

If we now use this estimate for σ we arrive at the *externally Studentized* residual:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}.$$

Typically, r_i and t_i give very similar values, however, the size of the discrepancy between them will depend on the influence of the i^{th} data point. Our formula for s_{-i} shows that a point with a large ordinary residual and a large leverage will produce a sizable difference between s and s_{-i} and thus between r_i and t_i.



Hypothesis test for outliers

20

Often, the externally Studentized residual will be a more sensitive detector of outliers.

We assume that the ϵ_i's are indeed normally distributed. Then each t_i is distributed according to a t-distribution with $n-p-1$ degrees of freedom under the assumption that the i^{th} data point does not suffer from a location shift.

$H_0 : \Delta_i = 0$, (this is not true for the internally Studentized residual, r_i, since, unlike t_i it cannot be written as a ratio of two independent quantities).

This provides us with a formal mechanism to test for the presence of location shift outliers.

In fact, the externally Studentized residual will also pick up scale shifts. Of course, we should recall our initial warning about the treatment of outlier detection as a formal statistical technique instead of a *simple diagnostic tool*.