Diagnostics

- Internally studentized residuals, PRESS residuals or externally studentized (case-deleted) residuals.
- Leverage.
- An individual point have large impact on $\hat{\mu}_i$. Diagnostic tool: DFFITS.
- An individual point may have large impact on $\hat{\mu}$ (the whole vector). Diagnostic tool: Cook's distance.
- An individual point may have large impact on $\hat{\beta}$. Diagnostic tool: DFBETAS.
- Modified Levene test for heteroscedasticity; see text.
- Breusch-Pagan test for heteroscedasticity; see text.
- Shapiro Wilk test for normality.
- Added variable plot.
- Pure error sum of squares F test.

Leverage

- Leverage i is h_{ii} — diagonal entry in hat matrix, H.
- $Var(\hat{\mu}_i) = h_{ii}$ and $Var(\hat{\epsilon}_i) = 1 - h_{ii}$ so $0 \leq h_{ii} \leq 1$.
- $trace(H) = p$ so the h_{ii} average to p/n.
- Rule of thumb. $h_{ii} > 2p/n$ is "large" leverage.
- Rule of thumb. $h_{ii} > 0.5$ is large, $h_{ii} > 0.2$ is moderately large.

DFFITS

Measure change in fitted value for case i after deleting case i:

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

- Any subscript (i) refers to a computation with case i deleted.
- Can be computed from externally deleted residual by multiplying by $\sqrt{h_{ii}/(1-h_{ii})}$. Thus can be computed without actually deleting case i and rerunning.
- Rule of thumb from text: look out for $|DFFITS| > 1$ in small to medium data sets or for $|DFFITS| > 2\sqrt{p/n}$ in large data sets.
- But I just examine the few largest values.

Cook's Distance

An individual point may have large impact on $\hat{\mu}$ (the whole vector):

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\text{MSE}}$$

- Can be computed without deleting case from

$$D_i = \frac{\hat{\epsilon}_i^2}{p\text{MSE}}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

- To judge size compare to $F_{p,n-p,0.90}$ (lower tail area is 10%, usually found as 1 over upper 10% point of $F_{n-p,p}$) and to $median(F_{p,n-p})$.
- Bigger than latter is quite serious.
- Smaller than former is good.
- Between is gray zone.

DFBETAS

- Intended to measure impact of deleting case i on $\hat{\beta}_k$.
- Defined by:

$$DFBETAS_{k(i)} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)}\left[(X^T X)^{-1}\right]_{kk}}}$$

- Same guidelines as DFFITS.
- Software not always set up to compute DFBETAS.

Tests for Homoscedasticity

- Modified Levene test:
  - Split data set into 2 parts on basis of covariates
  - Fit regressions in each part separately.
  - Do 2 sample t-test on mean absolute size of residuals.
- Breusch-Pagan test:
  - Regress squared fitted residual on covariate or covariates
  - Test for non-zero slope.

Tests of Distributional Assumptions

- Check assumption of Normality.
- Examine $Q - Q$ plot for straightness.
- Shapiro-Wilk test applied to residuals
- Or correlation test in Q-Q plot.

Pure Error Sum of Squares

- Sometimes for each (or at least sufficiently many) combination of covariates in a data set, there are several observations.
- Can do extra sum of squares F-test to see if our regression model is adequate.
- Suppose that $x_1, \ldots, x_K$ are the **distinct** rows of the design matrix
- Suppose we have n_1 observations for which the covariate values are those in x_1, n_2 observations with covariate pattern x_2 and so on. Of course $n_1 + \cdots + n_K = n$.
- We compare our final fitted model with a so-called saturated model by an extra sum of squares F-test.
- To be precise let α_1 be the mean value of Y when the covariate pattern is x_1, α_2 the mean corresponding to x_2 and so on.
- Relabel the n data points as $Y_{i,j}; j =, \ldots, n_i; i = 1, \ldots, K$
- Fit a one way ANOVA model to the $Y_{i,j}$.

Pure Error Sum of Squares

- Error sum of squares for this FULL model is

$$ESS_{FULL} = \sum_{i=1}^{K} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_{i,\cdot})^2$$

- This ESS is called the pure error sum of squares because we have not assumed any particular relation between the mean of Y and the covariate vector x.
- We form the F statistic for testing the overall quality of our model by computing the "lack of fit SS" as

$$ESS_{Restricted} - ESS_{FULL}$$

where the restricted model is the final model whose fit we are checking.

Example: plaster hardness

- 9 different covariate patterns: 3 levels of SAND and 3 levels of FIBRE.
- Two ways to compute pure error sum of squares:
  - Create new variable with 9 levels.
  - Fit a two way ANOVA with interactions.

DATA

0	0	1	61	34
0	0	1	63	16
15	0	2	67	36
15	0	2	69	19
		⋮		
30	50	9	74	48

SAS CODE

```
data plaster;
infile 'plaster1.dat';
input sand fibre combin hardness strength;
proc glm  data=plaster;
 model hardness = sand fibre;
run;
proc glm  data=plaster;
 class sand fibre;
 model hardness = sand | fibre ;
run;
proc glm  data=plaster;
 class combin;
 model hardness = combin;
run;
```

EDITED OUTPUT

Complete output

Source	DF	Sum of Squares	Mean Square	F	Pr > F
Model	2	167.41666667	83.70833333	11.53	0.0009
Error	15	108.86111111	7.25740741		
Total	17	276.27777778			

Source	DF	Sum of Squares	Mean Square	F	Pr > F
Model	8	202.77777778	25.34722222	3.10	0.0557
Error	9	73.50000000	8.16666667		
Total	17	276.27777777			

Source	DF	Sum of Squares	Mean Square	F	Pr > F
Model	8	202.77777778	25.34722222	3.10	0.0557
Error	9	73.50000000	8.16666667		
Total	17	276.27777778			

From the output we can put together a summary ANOVA table

Source	df	SS	MS	F	P
Model	2	167.417	83.708		
Lack of Fit	6	35.361	5.894	0.722	0.64
Pure Error	9	73.500	8.167		
Total (Corrected)	17	276.278			

- F statistic is $[(108.86111111 - 73.50000000)/6]/[8.16666667]$.
- P-value comes from the $F_{6,9}$ distribution.
- P-value not significant: no reason to reject final fitted model which was additive and linear in each of SAND and FIBRE.
- Notice that the Error SS are the same for the two-way ANOVA with interactions, which is the second model, and for the 1 way ANOVA.

- This test is not very powerful in general.
- More sensitive tests are available if you know how the model might break down.
- For instance, most realistic alternatives will be picked up more easily by checking for quadratic terms in a bivariate polynomial model; see earlier lectures.
- Notice that test for any effect of SAND and FIBRE carried out in the one way analysis of variance is not significant.
- This is an example of the lack of power found in many F-tests with large numbers of degrees of freedom in the numerator.
- If you can guess a reasonable functional form for the effect of the factors (either the additive two way model with no interactions or the even simpler multiple regression model which is the first model above) you will get a more sensitive test usually.

Added Variable Plots or partial regression plots

- Regress Y on some covariates X_1.
- Get Residuals.
- Regress other covariate X_2 on X_1.
- Get Residuals.
- Plot two sets of residuals against each other.

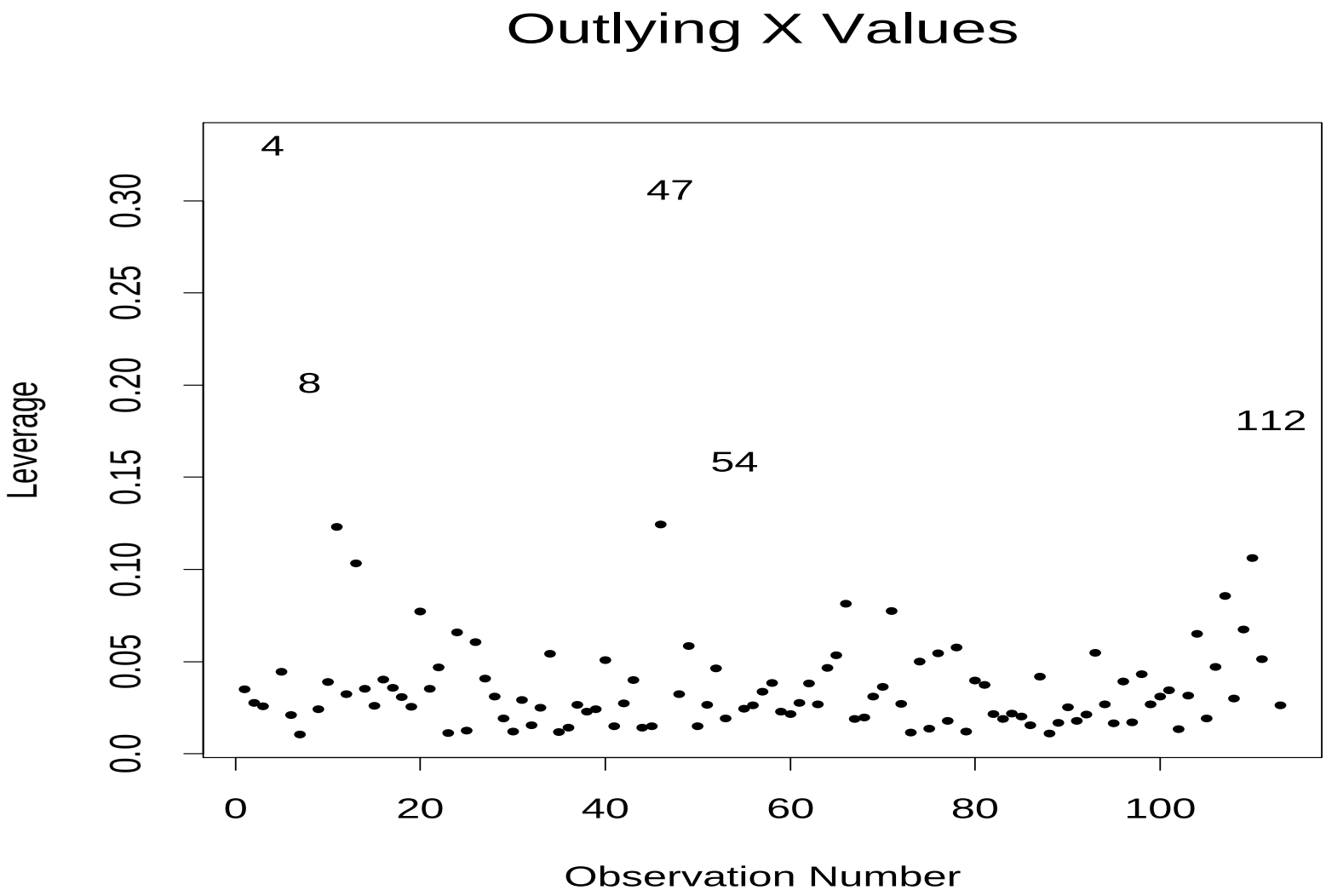
SENIC data example

Fit final selected model: covariates used are STAY, CULTURE, NURSES, NURSE.RATIO.

```
options pagesize=60 linesize=80;
data scenic;
 infile 'scenic.dat' firstobs=2;
 input Stay  Age Risk Culture Chest Beds
          School Region Census Nurses Facil;
 Nratio = Nurses / Census  ;
proc glm  data=scenic;
  model Risk = Culture Stay Nurses Nratio ;
  output out=scout P=Fitted PRESS=PRESS H=HAT
   RSTUDENT=EXTST R=RESID DFFITS=DFFITS COOKD=COOKD;
run ;
proc print data=scout;
```

Complete SAS Output is here.

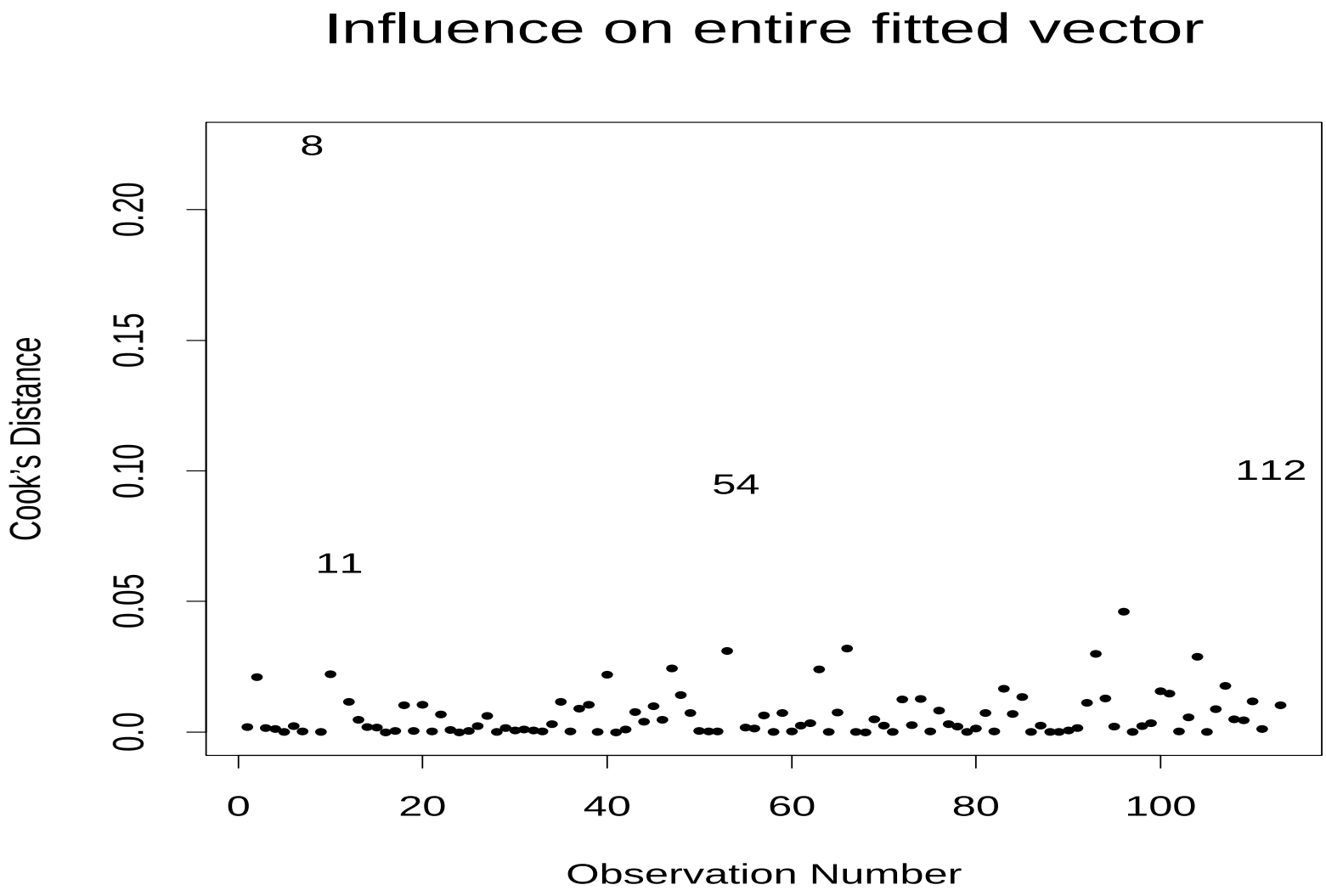
Index plot of leverages



Index plot of leverages: discussion

- Observations 4, 8, 47, 54 and 112 have leverages over 0.15
- Many more are over 10/113 — the suggested cut off.
- I prefer to plot the leverages and look at the largest few.
- Observations 4 and 47, in particular, have leverages over 0.3 and should be looked at.
- That means scientist thinks about those hospitals!

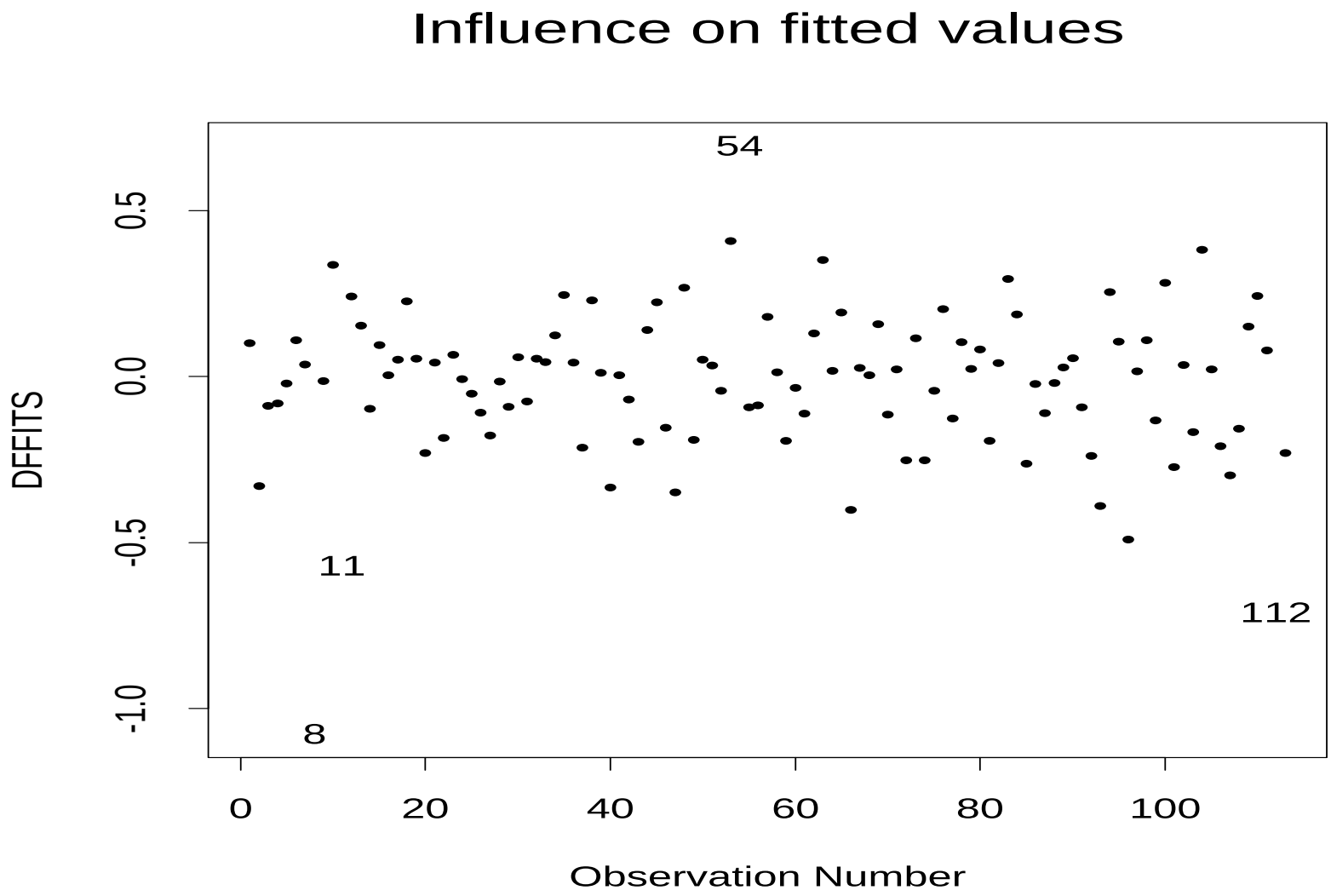
Influence measures: Cook's Distance



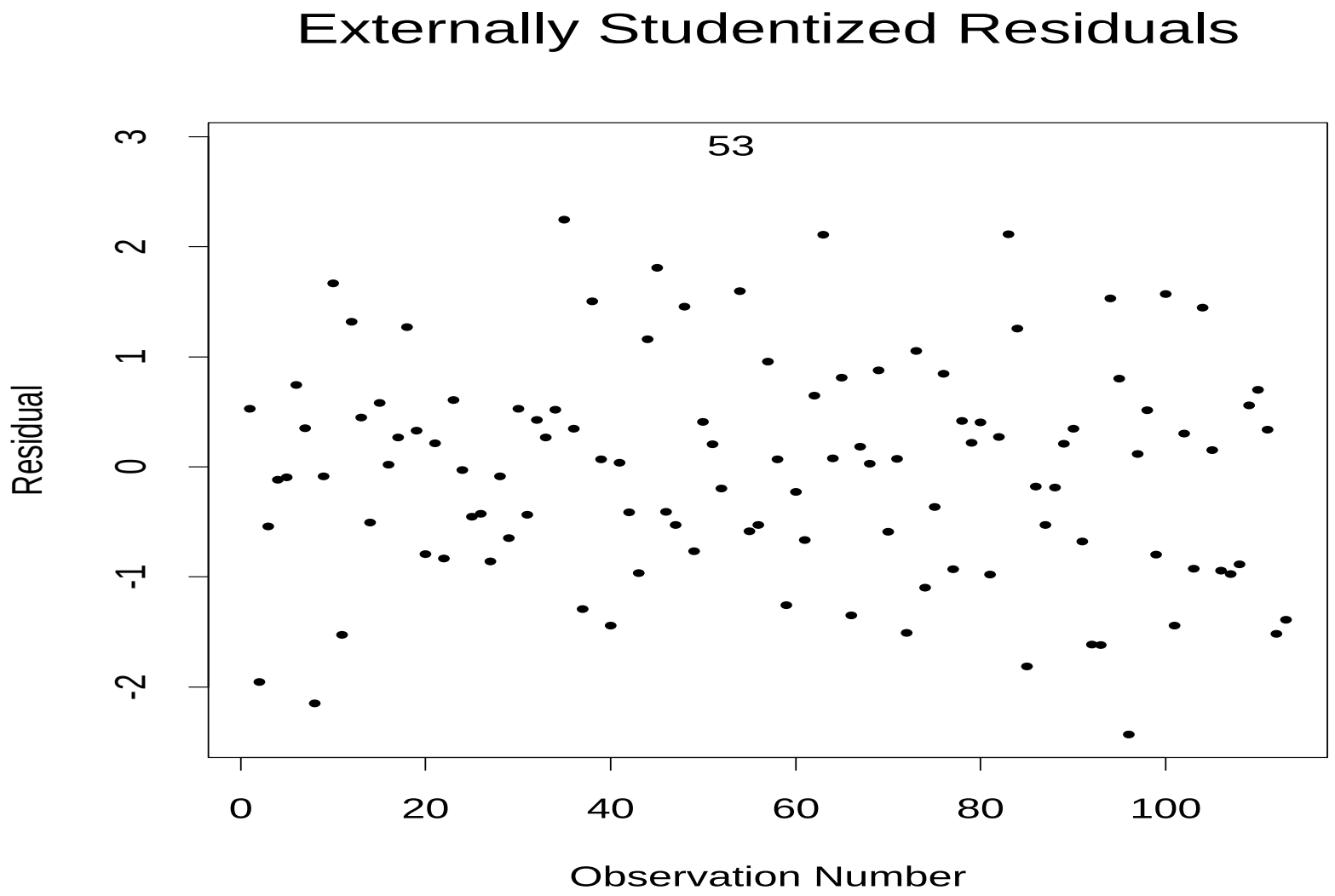
Cook's distance: discussion

- Observations 8, 11, 54 and 112 have values of D_i larger than 0.05.
- Of these, only observation 11 is new.
- Text recommends worrying only about observations for which D_i is larger than the tenth to twentieth percentile of the $F_{p,n-p}$ distribution.
- In this case those critical points are 0.3? and 0.46.
- None of the observations exceeds even the lowest of these numbers.

Influence measures: DFFITS



Case deleted residuals



Case deleted residuals: discussion

- Only observation 53 is added for our consideration,
- But with 113 residuals a value of 2.9 is not terribly unusual.

Examine observations highlighted by diagnostics

Here are the covariate values for observations 4, 8, 11, 47, 53, 54 and 112:

Observation	Culture	Stay	Nurses	Nratio	Risk
4	18.9	8.95	148	2.79	5.6
8	60.5	11.18	360	0.90	5.4
11	28.5	11.07	656	1.11	4.9
47	17.2	19.56	172	0.63	6.5
53	16.6	11.41	273	0.83	7.6
54	52.4	12.07	76	0.66	7.8
112	26.4	17.94	407	0.51	5.9
Mean	15.8	9.65	173	0.95	
SD	10.2	1.91	139	0.11	

- Observation 4 has a quite unusual value of Nurse.Ratio – a lot of nurses
- Observation 47 has quite a high average Stay for patients.
- The others are harder to interpret but 4 and 47 are the most leveraged observations.
- In summary it appears that several observations exert excess influence on the fitting process.
- As a final method of judging whether or not our fit was unduly influenced by these observations I fit the model again in SAS but removing observations number 4, 8, and 47.

```
                 Sum of    Mean
Source    DF   Squares   Square    F    Pr > F
Model      4  100.4617 25.1154 28.21 0.0001
Error    105   93.4950  0.8904
Total    109  193.9567
 R-Square     C.V.    Root MSE    RISK Mean
 0.517959  21.87080 0.9436255   4.3145455
                  T for H0: Pr > |T|   SE
Parameter      Est  Par=0              Est
INTERCEPT -.15118  -0.21   0.8349   0.7237
CULTURE    0.05686   5.28   0.0001   0.0108
STAY       0.27735   4.18   0.0001   0.0663
NURSES     0.00167   2.30   0.0232   0.0007
NRATIO     0.70245   1.92   0.0578   0.3661
```

Compare these results to the corresponding parts of the same code applied to the full data set.

```
Dependent Variable: RISK
                 Sum of    Mean
Source   DF     Squares  Square     F    Pr > F
Model     4   103.6905 25.9226 28.66 0.0001
Error   108    97.6893  0.9045
Total   112   201.3798
  R-Square    C.V.    Root MSE  RISK Mean
  0.514900 21.83920 0.9510681 4.3548673
                     T for H0: Pr > |T|  SE
Parameter  Estimate  Par=0              of Est
INTERCEPT  -.083138 -0.14    0.8917   0.6092
CULTURE    0.048249  5.03    0.0001   0.0096
STAY       0.276744  5.04    0.0001   0.0549
NURSES     0.001587  2.26    0.0258   0.0007
NRATIO     0.769487  2.57    0.0115   0.2994
```

Summary: differences seem minor; little harm in sticking to model fitted earlier.

Making an Added variable plot: example

- For SENIC data to assess influence of facilities.
- Regress RISK on STAY, CULTURE, NURSES, NURSE.RATIO. Get residuals.
- Regress FACILITIES on STAY, CULTURE, NURSES, NURSE.RATIO. Get residuals.
- Plot residuals against each other. Look for patterns.

Added Variable Plot for FACILITIES