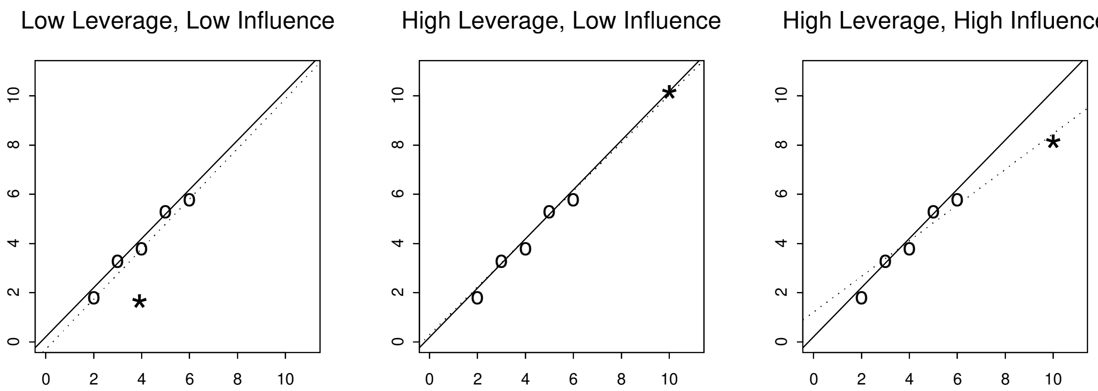
36-401 Class Notes and Exercise
Regression Criticism II: Outliers, Influence & Leverage

Outliers: Influence and Leverage (Potential)

Some General Ideas

The notions of "outlying" or "influential" data are deliberately somewhat vague notions in Statistics. In linear models we usually deal with three related notions:

- A data point $(y_i, x_{i1}, \ldots, x_{ik})$ is *outlying* if it is far from the remainder of the data.
- A data point has high *leverage* if parameter estimates or fitted values change a great deal when the data point is changed by a small amount—this is an inherently differential-calculus notion.
- A data point is *influential* if removing it from the data set changes the parameter estimates or fitted values a great deal.



Outlying points ($*$) illustrating varying degrees of leverage and influence. The dotted line is the linear fit without the oulier, the solid line is the fit with the outlier.

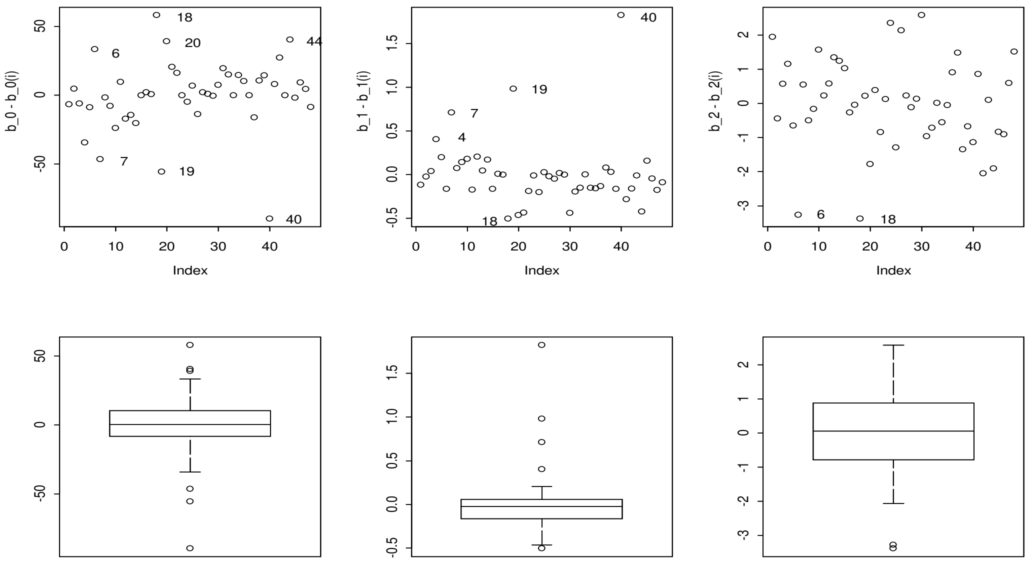
From the figure, we can deduce sevearal important heuristics for linear regression:

- Outlyingness in X seems to drive *leverage*, but may or may not be related to lack of model fit.
- Outlyingness in Y is a clear indication of *lack of model fit*; and combined with high leverage seems to drive influence: *Influence* $=$ *Leverage* $\times$ *Outlyingness*.
- To measure *influence*, compare the regression with and without the data point.

- *Dealing with Influential Cases*
    - ◇ Check for clerical or experimental mistakes
    - ◇ Check non-included covariates that are "special" for this case
    - ◇ Consider transformations if the influential cases come from a skewed tail
    - ◇ Consider robust regression, especially if the errors appear to be heavy tailed
    - ◇ Consider reporting results with and without outliers; think about whether the outlier is beyond a region of linearity
    - ◇ Cautiously consider throwing out gross outliers that are probably mistakes, even though we have no evidence of such.

- *DFBETAS: Comparing b_k (using all observations) with $b)k(i)$ (with observation i omitted)?*

    Example:
    In the Fuel consumption data set we had 48 observations (after excluding Hawaii and Alaska) and we fitted the model $\mathit{Fuel} = \beta_0 + \beta_1 \mathit{Dlic} + \beta_2 \mathit{Tax} + \epsilon$. Here are plots of $b_0 - b_{0(i)}$ (left), $b_1 - b_{1(i)}$ (middle) and $b_2 - b_{2(i)}$ (right), after leaving out the i^{th} observation:



Plots of $b_k - b_k(i)$ for `lm(formula = Fuel ~ Dlic + Tax, data = fuel48)`

A slighly better tool is

$$\text{DFBETAS}_{ik} = \frac{b_k - b_{k(i)}}{s_{e(i)}/\sqrt{\text{RSS}_k}}$$

Note that we have divided by an estimate of SE($b_{k(i)}$), so we can use our intuition about $N(0,1)$ rv's to see unusual values right in the numbers: Observation i is unusual, if

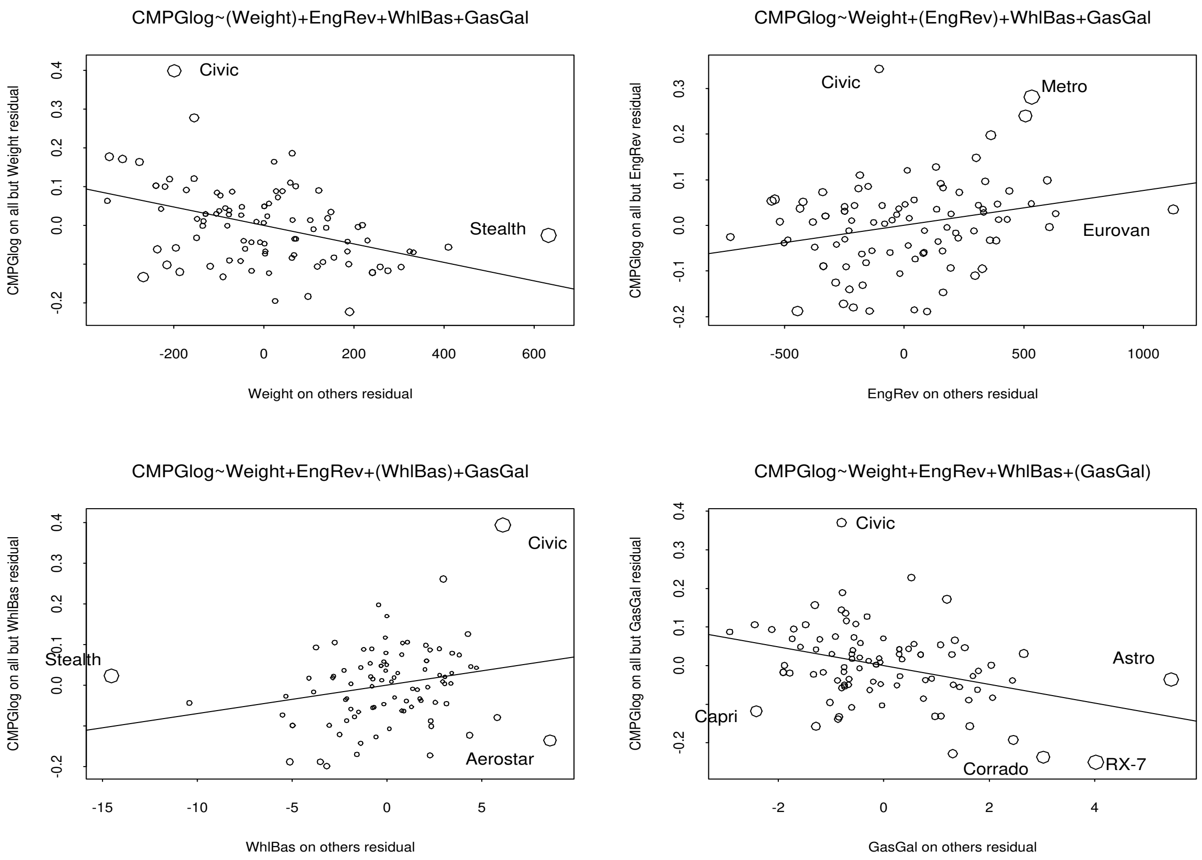
- ◇ $|\text{DFBETAS}_{ik}| > 2/\sqrt{n}$; or
- ◇ severe outliers by box-plot (beyond 3 IQR's from from the 1st or 3rd quantile); or
- ◇ separated from the rest of the DFBETAS for the other cases by a large gap.

A proportional leverage (influence) plot is a partial regression plot for X_k with point size proportional to the DFBETAS$_{ik}$'s for each case, i.

Consider the car model

log(CityMPG) = β_0 + Wght β_1 + WhlBas β_2 + EngRev β_3 + GasGal β_4 + ϵ

```
cars$CMPGlog_log(cars$CitMPG)
rslt_lm(CMPGlog~Weight+WhlBas+EngRev+GasGal, cars)
summary(rslt)

Coefficients:  Value Std. Error  t value Pr(>|t|)
(Intercept)   3.3085   0.2542    13.0172   0.0000
     Weight  -0.0002   0.0001    -4.2597   0.0001
     WhlBas   0.0070   0.0030     2.3607   0.0205
     EngRev   0.0001   0.0000     2.5571   0.0123
     GasGal  -0.0243   0.0069    -3.5374   0.0006

Residual standard error: 0.09501 on 88 degrees of freedom
Multiple R-Squared: 0.8277

source("partreg.q")

par(mfrow=c(2,2))
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="Weight", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="EngRev", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="WhlBas", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
partreg(xlst=c("Weight","EngRev","WhlBas","GasGal"), yname="CMPGlog",
   outname="GasGal", df=cars, prop=T, id=T, tcex=0.7, cex=0.8)
```



- *Diagonal entries of hat matrix h_{ii} : measures of Leverage (Potential)*

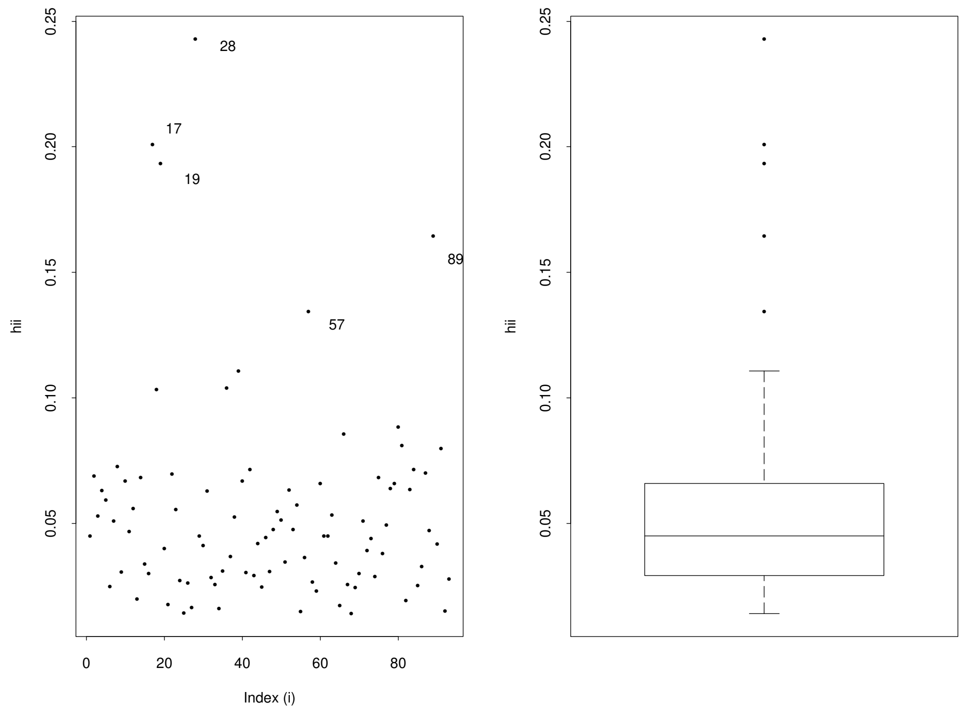
$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] = [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

The leverage of the ith case is h_{ii}, the ith diagonal element of $\mathbf{H}$. $1/n <= h_{ii} < 1$ and since the ith row of $\mathbf{H}$ defines the linear combination of the Y_i's that make $\hat{Y_i}$, it makes sense to think that large h_{ii} indicates that case i has the potential to have a large overall effect, i.e. $\hat{Y_i}$ is approximately Y_i and the regression line is "forced" to go through (X_i, Y_i).

Criteria for "high leverage"

- ◇ $h_{ii} > 2K/n$ (since $\bar{h_{ii}} = K/n$)
- ◇ $\max(h_{ii}) > 0.5$; $0.2 < \max(h_{ii}) < 0.5$ is risky
- ◇ Make a boxplot or stem and leaf plot of the h_{ii}'s and look for outliers or well-separated h_{ii}'s.

```
> hii _ lm.hat(rslt)            # Does the linear algebra for you!
> c(sum(hii),mean(hii),2*mean(hii))
[1] 5.00000000 0.05376344 0.10752688
> par(mfrow=c(1,2))
> plot(hii,xlab="Index (i)",ylab="hii")
> identify(1:length(hii),hii)
[1] 57 89 19 17 28
> boxplot(hii,style.bxp="old",ylab="hii")
```



- *Standardized and Studentized Residuals*

  The variance-covariance matrix of the residuals, e_i's, is $\sigma^2(\mathbf{I} - \mathbf{H})$. We can at least correct for the unequal variances by using *standardized residuals*:

$$z_i = \frac{e_i}{s_e\sqrt{1 - h_{ii}}}$$

  A better version is the *studentized residual*, t_i, which replaces s_e with $s_{e(i)}$, the estimated residual s.d. with case i removed. These each follow a t-distribution. To correct for multiple comparisons we can say that a case residual is an outlier if t_i is outside of the 100($1 - \alpha/n$)% confidence limits for a t-distribution with n-K df.

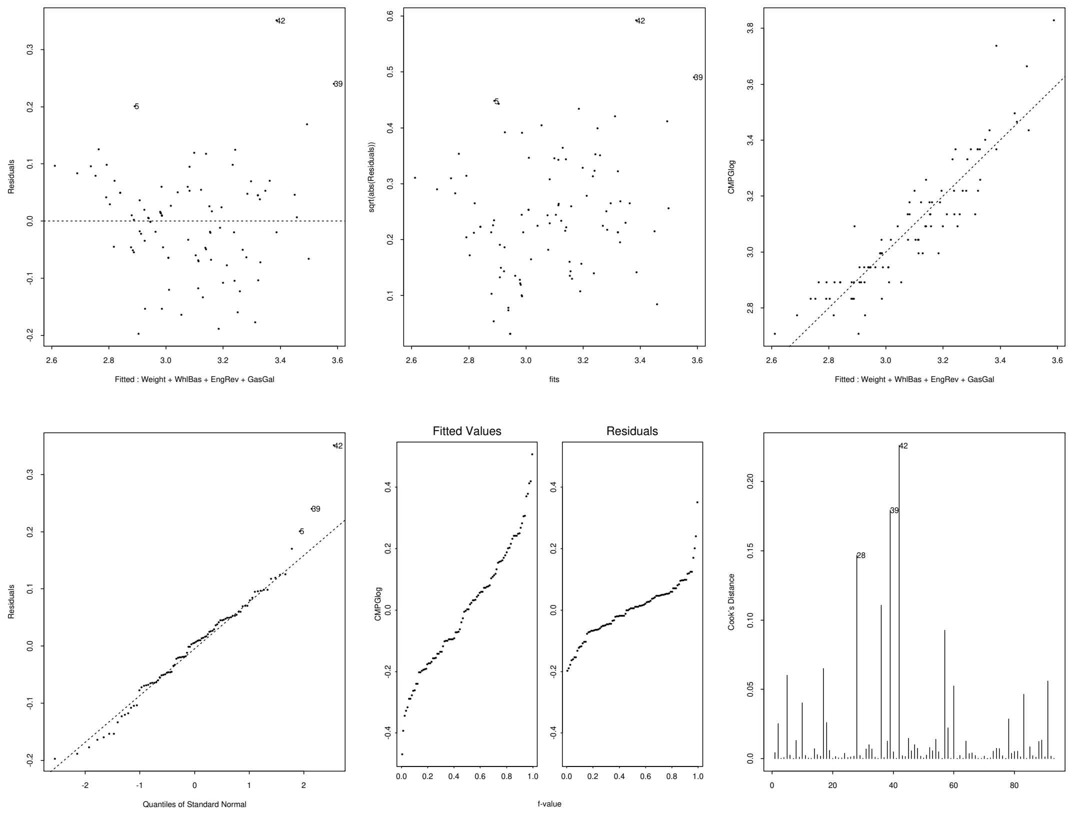
  [Example in the breakout exercise]

- *Cook's Distances D_i*

$$D_i = \frac{z_i^2 h_{ii}}{K(1 - h_{ii})}$$

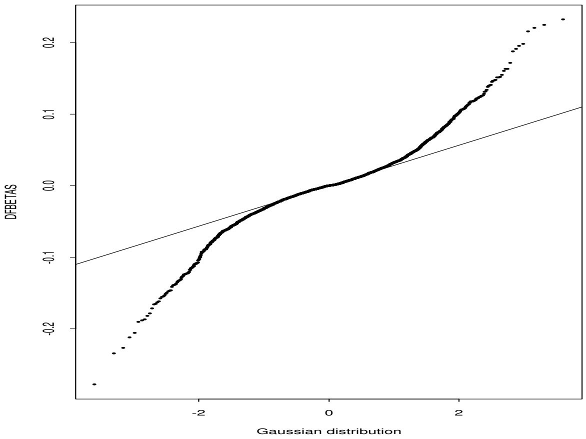
measures the *overall* influence of the ith case on all regression coefficients. Cases are considered influential if $D_i > 4/n$.

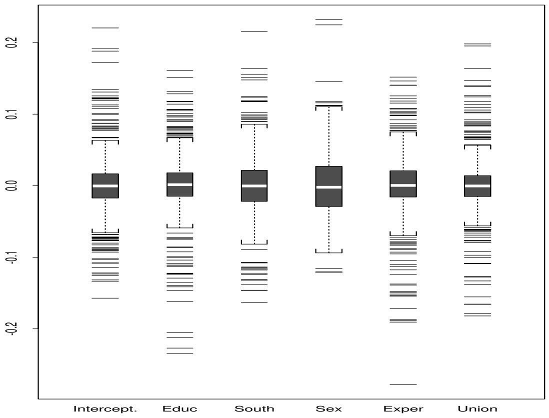
Cook's D_i is basically a summary of $DFBETAS_{ki}$ for all k.



Breakout Session

```
rslt_lm(logWage~Educ+South+Sex+Exper+Union,wages)
source("influence.q")
inf_influence(rslt)
names(inf) # "standres" "studres" "dfbetas" "dffit" "dffits" "hat" "cooksd"
nrow(wages) # 534

par(mfrow=c(1,2))
qqn(inf$dfbetas, ylab="DFBETAS", main="")
boxplot(data.frame(inf$dfbetas))
```

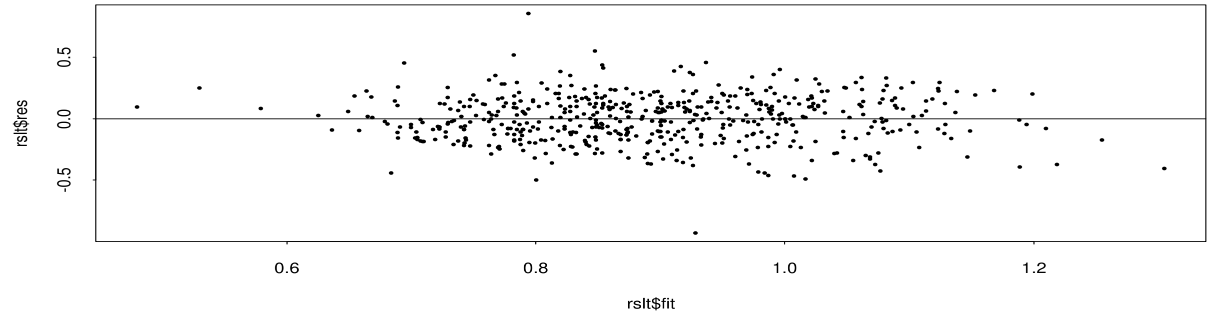


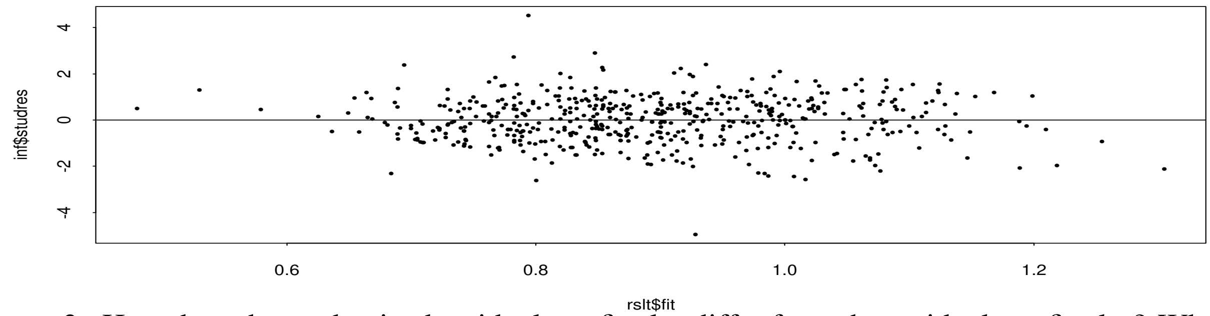



```
d_inf$dfbetas
Sel_apply(d,1,function(x){max(abs(x))>0.2})
print(round(d[Sel,],3))
#     (Intercept)   Educ  South    Sex  Exper  Union
# 108       0.191 -0.234  0.118 -0.116  0.140 -0.079
# 171       0.123 -0.048 -0.132  0.225 -0.278 -0.033
# 200      -0.123  0.063  0.163  0.232 -0.110  0.163
# 469       0.220 -0.227  0.011  0.072 -0.172 -0.166
# 478       0.172 -0.206  0.034  0.106 -0.150  0.074
# 482       0.188 -0.212  0.030  0.106 -0.188  0.076
# 495      -0.115  0.132  0.215  0.145 -0.083 -0.003

cbind(1:534,inf$hat)[inf$hat>0.025,]
#      [,1]       [,2]
# [1,]   63 0.04113022
# [2,]   76 0.02951064
# [3,]  108 0.02742524
# [4,]  347 0.03668933
# [5,]  351 0.04624977
# [6,]  493 0.02633355
# [7,]  500 0.02556922

cbind(1:534,round(inf$cooksd,4))[inf$cooksd>0.015,]
#     [,1]   [,2]
#  15   15 0.0154
# 108  108 0.0266
# 171  171 0.0297
# 200  200 0.0241
# 414  414 0.0220
# 469  469 0.0189
# 495  495 0.0183
```

1. What cutoff values are useful for the influence statistics used above? How do you interpret the graph and values shown? What would you do next?

```
plot(rslt$fit, rslt$res)
plot(rslt$fit, inf$studres)
```





2. How does the studentized residual vs. fit plot differ from the residual vs. fit plot? When might they be more different?