1/28/2010 36-402/608 ADA-II H. Seltman
Handout #6: Outliers

1. These tools should be used to evaluate models in conjunction with residual vs. fit, residual vs. x, and residual quantile-normal plots. It's best to use:

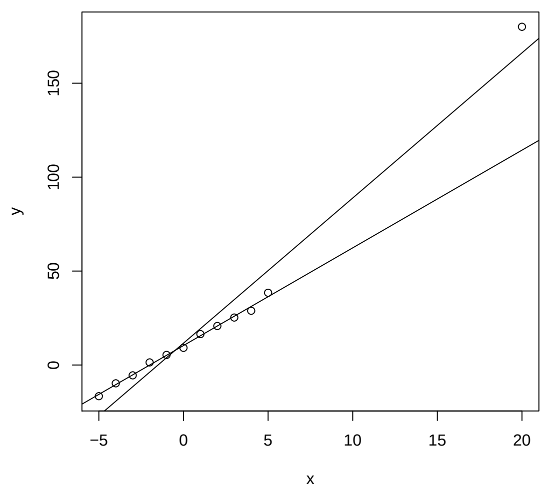
   `plot(rstudent(lm.result)~fitted(lm.result));abline(h=0)`.

2. High leverage points

   (a) Definition: "X" points that are far from other X points have the *potential* to have an unduly large effect on the "response surface". Easy to see in simple regression; harder to detect in multiple regression.

   (b) Example

```
x = c(seq(-5,5), 20)
y = 10 + 5*x + rnorm(length(x),0,2)
y[length(y)] = y[length(y)]+50
plot(x,y); abline(lm(y~x)); abline(lm(y[-12]~x[-12]))
dev.copy(pdf, "HO5A.pdf");dev.off()
```




```
summary(lm(y~x))
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)  11.6132     2.8776   4.036  0.00238
#x             7.7267     0.4414  17.505 7.86e-09
#Residual standard error: 9.637 on 10 degrees of freedom
#Multiple R-Squared: 0.9684,    Adjusted R-squared: 0.9652
```

```
summary(lm(y~x, subset=x<10))
#             Estimate Std. Error t value Pr(>|t|)
#(Intercept)  10.3500     0.4013   25.79 9.55e-10
#x             5.2003     0.1269   40.97 1.53e-11
#Residual standard error: 1.331 on 9 degrees of freedom
#Multiple R-Squared: 0.9947,    Adjusted R-squared: 0.9941
```

(c) Theory

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}},\ \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \Rightarrow \hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \quad \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

$$E(\hat{\boldsymbol{\epsilon}}) = \mathbf{0} \quad \mathrm{Var}(\hat{\boldsymbol{\epsilon}}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

$$\mathrm{Var}(\hat{\boldsymbol{\epsilon}}_i) = \sigma^2(1 - h_{ii})$$

Therefore, if `influence(lm(y~x))$hat[i]` is near 1 instead of near 0, then the residual for the i$^{\mathrm{th}}$ data point will be near zero, and the prediction ($\hat{Y}_i$) will be near Y_i. In other words, the regression line will be pulled near Y_i, so Y_i has "leverage". If removing Y_i doesn't change the regression line, Y_i is not actually "influential", just potentially so. A worrisomely large hat value is 2p/n.

(d) In R: `summary(influence.measures(lm(y~x)))`

```
Potentially influential observations of
         lm(formula = y ~ x) :
   dfb.1_  dfb.x   dffit   cov.r   cook.d  hat
12  3.18_* 41.44_* 43.82_*  0.00_* 18.32_*  0.79_*
```

3. Influential points: Y outliers especially when they have high leverage

(a) Cook's Distance: A (standardized) measure of how far the regression surface moves when observation i is eliminated from the data. The numerator is

$$\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2$$

where $\hat{Y}_j$ is the predicted Y_j when observation i is included and $\hat{Y}_{j(i)}$ is the predicted Y_j when observation i is excluded. Equivalently this is a measure of how far the whole set of parameter estimates move when point i is dropped. The rule-of-thumb is to worry when $D_i > 1$.

(b) DFFITS is similar to Cook's distance, but focuses on the effect of removing point i on the prediction for point i alone. Common cutoff's are 2 or $2\sqrt{p/n}$.

(c) DFBETAS is an n by p matrix with information about how much each coefficient estimate will change (in some standardized form) when point i is removed. Common cutoff's are 2 or $\sqrt{2/n}$.

(d) Covariance ratio (cov.r) is a measure of how much the overall uncertainty in the coefficient estimates changes when point i is removed. Values more than $3p/n$ from 1 are worrisome.

4. influencePlot() in package “car” is a nice interactive plot of (studentized) residuals vs. leverage (hat) with point size proportional to Cook’s D, and with interactive labeling of outliers. It returns the outlier row numbers.

5. Sleuth Algorithm

   (a) Start with residual plots, and proceed to influence analysis if anything looks amiss.

   (b) If deleting the suspect case(s) does not affect the conclusions, keep them (but check how they are different).

   (c) If deleting changes things, eliminate the case if there is a very good, clear reason to suspect that the case belongs to a different population or is a mistaken measurement.

   (d) If the case(s) is not clearly from a different population, but is far from other cases in the x directions, omit the case and, as usual, be careful about extrapolating outside the (remaining) x range of the data.

   (e) If the suspect case(s) is not an x outlier, no good conclusions can be made. Perhaps report results with and without the case.

6. Breakout and Discussion