## Stat 302 – Day 26
### Outliers, Leverage, and Influential Observations

### Last Time: Transformations
- Transformations can be considered if don’t meet basic model assumptions (LNE)
  - Normality less important with large sample sizes except for prediction intervals
  - Can transform one or both variables
- Back-transform prediction intervals before interpreting
  - With confidence interval, a log transformation becomes about the median response
  - With slope, a log transformation becomes about a multiplicative change in median response
- Interpret $R^2$ in terms of transformed response
  - So can’t compare $R^2$ across models if one has transformed response

(a) Recreate a “fitted line plot” and label Palm Beach.

**JMP:** Use Analyze > Fit Model. Click on an observation in the scatterplot (it will turn black, the others grey), the row will highlight in the Data window. In the Data window, choose Rows > Label. Then press Escape to unhighlight the row. Click on the County column to highlight it (only) and choose Cols > Label/Unlabel.

How can we measure just how large/unusual Palm Beach’s residual is? There are several methods for scaling the residuals.

**Definition:** The **standardized residuals** divide the residuals by their standard deviation. Consequently these have mean zero and standard deviation 1. Values larger than 2 or 3 are considered unusual.

**JMP:** Use the Response pull-down menu to select Save Columns > Studentized Residuals (this is actually an incorrect naming convention used by JMP so just go with it)

(b) Create a plot of the standardized residuals vs. *Bush* votes. The patterns will be the same but this graph can help you decide whether a residual is “significantly” large. Identify and interpret Palm Beach’s standardized residual.

### Definition: An observation is **influential** if removing it from the data set substantially changes the slope or the intercept or the correlation coefficient and $R^2$ values.

### Definition: Observations that are extreme in the $x$-direction are said to have the largest potential for influence. We measure the (standardized) distance from $\bar{x}$ using leverage or “hat” values. The leverage of the $i^{th}$ observation: $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}$. Note: $SE(\hat{y})$ for CI = $s\sqrt{\bar{h}}$

Many software packages will flag an observation with high leverage to remind you to consider its influence on the regression model. One nice result is $\Sigma h_i = 2$ and so $\bar{h} = 2/n$. Observations larger than $4/n$ or $6/n$ are considered far enough from the rest of the data to be considered high leverage and should be investigated. Another guideline is $h_i$ values exceeding .5 indicate high leverage and those
between .2 and .5 indicate moderate leverage. Best approach: examine a graph of the leverage values and look for observations that are substantially further away from the rest (relative) unusualness.

(c) Which observation do you think will have more leverage, Palm Beach or Dade County?

(d) Save and create a histogram or dotplot of the leverage values. Which observation(s) stand out? JMP: Use the Response pull-down menu to select Save Columns > Hats.

But keep in mind that not all high leverage points are indeed influential. So what we really need is a way to measure whether individual observations have a substantial influence on the regression equation. We will focus on changes to the regression coefficients (and consequently the predicted values).

One way to measure the influence on the predicted values is to compare the fitted values with the observation in the data set and the fitted values without the observation in the data set $\left( \hat{y}_i - \hat{y}_{i(i)} \right)$.

To standardize this difference, we will divide by $se(\hat{y}) = s(i)\sqrt{h_i}$ where $s(i)$ is the estimate of $\sigma$ with the potentially influential observation removed.

$$DFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{s(i)\sqrt{h_i}}$$

Thus, $DFITS_i$ is the number of standard deviations that the fitted value $\hat{y}_i$ changes if observation $i$ is removed.

Similarly, **Cook’s Distance (Cook’s D Statistic)** is an aggregated measure of the effect of removing the $i^{th}$ observation on all of the predicted values.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{2\hat{s}^2} \left( \frac{h_i}{1-h_i} \right)$$

This can be interpreted as the squared Euclidean distance that the vector of fitted values moves when the $i^{th}$ observation is deleted. For now, we will consider an observation moderately influential if $D_i > .5$ and highly influential if $D_i > 1$.

(e) Store the Cook’s D statistic. Are either Palm Beach or Dade County considered highly influential? JMP: Use the Response pull-down menu to select Save Columns > Cooks D Influence.

(f) Based on the second formula, what needs to be true for an observation to be influential? (BTW, why is this formula considered “computationally advantageous”?)

**Notes:**
- These are measures of individual influence, could still be a group or cluster with influence. Are “multiple case” detection methods as well.
- Can still investigate effect on individual research questions.
- Can also measure influence on the regression coefficients (Delta BETAS). (R: dfbeta)
Example 2: The worksheet stateSAT.txt contains data on the average SAT scores by state. Researchers have tried to explain the state by state differences in scores (e.g., Powell and Steelman, (1984), “Variations in SAT Performance: Meaningful or Misleading?” Harvard Educational Review 54(4)). Column 2 is the average SAT scores, along with six variables that may be associated with the SAT differences among states: percentage of the total eligible students who took the exam, median income of families of test takers, average number of years that the test takers had formal studies in social studies, natural sciences, humanities, percentage of test takers who attended public secondary schools, total state expenditure on secondary schools (dollars per student), and median percentile ranking of the test takers within their secondary school classes.

(a) Examine a fitted line plot of the SAT and income. Do any states stand out as not following the same pattern of the rest of the states? Identify the states by name and describe their unusual behavior. Do you think the regression line would change if these states were removed? Explain.

(b) Which states have the largest residuals? Are they positive or negative? What does that imply in this context?

(c) Do any states appear to have large leverage?
   - Produce a graph of the leverage values (or resids vs. leverage). Do any points stand out? Which is the most extreme? Does its \( h_i \) value exceed \( 4/n \)? Any others?

(d) Do any states appear to have large influence?
   - Produce a graph of the Cook’s \( D_i \) values. Do any points stand out? Do any exceed 0.5 or 1?

(e) Would we have any justification for removing this observation from the data set?

(e) Where could you put an observation in this dataset, so that it is not an outlier in either the x direction or the y direction but is an outlier in the regression? Do you think this observation will have much influence on the regression line?
Dealing with Influential Observations
Finding the causes of unusual observations and deciding what to do with them is a more difficult issue.

- Can’t just throw them away!
- Is it truly influential? If not, can proceed anyway
- Review data collection
  - Does the observation come from a different population? If so, can remove.
  - Is it a typo?
- If it is the only observation in a region, can restrict the region of interest (with comment)
- Conservative approach: fit model with and without unusual observations
- Robust or resistant regression procedures (e.g., median-median line, rlm)
- If possible: collect more data

*Unusual observations can sometimes tell the analyst as much (or more) about the underlying random process as other observations.*

(d) Compare the above results to either JMP’s robust equation or R’s iterated re-weighted least squares which downweights observations with large residuals.

JMP: Analyze > Fit Y by X; pull-down menu > Robust > Fit Robust

R:
> library("MASS")
> summary(rlm(SAT~income))

What has changed and why?

(f) But our real question here is how to rate the educational quality of the states based on the SAT scores. Is it reasonable to just take the states with the highest average SAT scores and call them best? Why or why not?

(g) How might we take some of these other issues into account?