Influence in Regression

- Outliers are data with an extreme value of the response variable (Y).
- Leverage points are data with an extreme value of the predictor variable (X).
- Some combination of extreme Y (outlier) and extreme X (leverage) makes a data point influential.
- An influential data point: removing the data point substantially changes the regression results.

How do we define “substantial”? 

Influence: Cook’s Distance

\[ D_i = \frac{\sum_{j=1}^{n} (\hat{y}_{ij} - \hat{y}_j)^2}{p s_{\epsilon}^2} \]

where
\[ D_i = \frac{e_i^2 h_{ii}}{p s_{\epsilon}^2 (1 - h_{ii})^2} = \frac{i s r_i^2}{p (1 - h_{ii})} \]

More convenient form:

\[ D_i = \left( \frac{e_i^2}{p s_{\epsilon}^2} \right) \frac{1}{(1 - h_{ii})} \]

The Cook’s Distance is considered significant if it is bigger than about \( \frac{4}{n} \) (alternately, use the 50th percentile of the \( F_{p,n-p} \) distribution).

Measuring Influence

- The Cook’s Distance is a measure of influence, but it is not a statistical test.
  - Outliers are not necessarily influential, and influential points are not necessarily outliers.
  - We don’t remove or adjust highly influential points.
  - Our goal is to identify influential points.
  - We worry about fragility: our conclusions depend only on 1 or 2 data points.

More Influence Measures
- For each $\beta_k$ of interest, find its estimate with and without the $i^{th}$ data point
  \[ DFBETA_{k,i} = \frac{b_k - b_k^{(i)}}{SE(b_k^{(i)})} \]
  Considered significant if DFBETA is bigger than about $2/\sqrt{n}$
- A measure similar to the Cook’s Distance is
  \[ DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i(0)}}{s_{e(0)} \sqrt{h_i}} = esr \frac{h_i}{\sqrt{1 - h_i}} \]
- Also, Mahalanobis Distance (we won’t discuss)

Review of Influence Measures

<table>
<thead>
<tr>
<th>Metric</th>
<th>Equation</th>
<th>Small Sample Criterion</th>
<th>Large Sample Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cook’s Distance</td>
<td>$D_i = \frac{(h_i - h_{i(0)})}{p(1 - h_{i(0)})}$</td>
<td>1</td>
<td>$\frac{4}{p/n}$ F.INV(0.5)</td>
</tr>
<tr>
<td>Difference in Beta</td>
<td>$DFBETA_{A,i} = \frac{b_k - b_k^{(i)}}{SE(b_k^{(i)})}$</td>
<td>1</td>
<td>$\sqrt{4/n}$</td>
</tr>
<tr>
<td>Difference in Fit</td>
<td>$DFFITS_i = \frac{\hat{y}<em>i - \hat{y}</em>{i(0)}}{s_{e(0)} \sqrt{h_i}} = esr \frac{h_i}{\sqrt{1 - h_i}}$</td>
<td>1</td>
<td>$\sqrt{4p/n}$</td>
</tr>
</tbody>
</table>

"Small" means about $n = 20$ or less

Multiple Regression
- When regressing on two or more predictor variables, it is best to let a statistical software package calculate quantities like $ist_i, esr_i, h_{ii}, D_i, DFBETA_{k,i}$ etc.
- Additionally, with multiple regression we have to worry about correlations between predictor variables
  - We’ll cover this later

Experimental Design
- An important goal of Design of Experiments (DoE) is to equalize the leverage of every point during multiple regression
  - We want to make $h_{ii} = \frac{p}{n}$ for every $i$
  - More on DoE later

Conclusions
- When regressing, outliers and high leverage data points are important to consider, but it is influential data that matters most
- When regressing, calculate for every data point: $ist_i, esr_i, h_{ii}, D_i, DFBETA_{k,i}$ etc.
- Use the Williams graph and graphs of the Cook’s Distance to get a feel for influence
- Consider deleting or altering outliers only if they are influential
- If your results are fragile, consider collecting more data to reduce the influence of the few data points that make your results fragile

Lecture 22: What have we learned?
- Define influence
- Name several metrics of influence
- Explain what is meant by a “fragile” regression
- How does a measure of influence affect the way we approach outliers?