Example: 11 Regions in the UK

<table>
<thead>
<tr>
<th>Region</th>
<th>Alcohol</th>
<th>Tobacco</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. North</td>
<td>6.47</td>
<td>4.03</td>
</tr>
<tr>
<td>2. Yorkshire</td>
<td>6.13</td>
<td>3.76</td>
</tr>
<tr>
<td>3. Northeast</td>
<td>6.19</td>
<td>3.77</td>
</tr>
<tr>
<td>4. East Midlands</td>
<td>4.89</td>
<td>3.34</td>
</tr>
<tr>
<td>5. West Midlands</td>
<td>5.63</td>
<td>3.47</td>
</tr>
<tr>
<td>6. East Anglia</td>
<td>4.52</td>
<td>2.92</td>
</tr>
<tr>
<td>7. Southeast</td>
<td>5.89</td>
<td>3.2</td>
</tr>
<tr>
<td>8. Southwest</td>
<td>4.79</td>
<td>2.71</td>
</tr>
<tr>
<td>9. Wales</td>
<td>5.27</td>
<td>3.53</td>
</tr>
<tr>
<td>10. Scotland</td>
<td>6.08</td>
<td>4.51</td>
</tr>
<tr>
<td>11. Northern Ireland</td>
<td>4.02</td>
<td>4.56</td>
</tr>
</tbody>
</table>
> BritRegions <- lm(Alcohol~Tobacco, data=BritishRegions)
> summary(BritRegions)

Coefficients:
    Estimate Std. Error t value Pr(>|t|)  
(Intercept)  4.3512     1.6067   2.708   0.0241 *  
Tobacco       0.3019     0.4388   0.688   0.5087    

Residual standard error: 0.8196 on 9 degrees of freedom  
Multiple R-squared:  0.04998, Adjusted R-squared: -0.05557  
F-statistic: 0.4735 on 1 and 9 DF,  p-value: 0.5087

For leverage, flag values with hi > 4/11 = .363

None are flagged b/c Scotland and N. Ireland are both far from the center.  Southwest far from center at low end.

> options(digits=3)
> hatvalues(BritRegions)

    1      2      3      4      5      6      7      8      9     10  11
0.1395  0.0967  0.0975  0.1131  0.0972  0.2306  0.1410  0.3273  0.0931  0.3188  0.3451

Southwest Scotland N. Ireland

Standardized residuals: Flag if absolute value > 2; No. Ireland is moderate

> rstandard(BritRegions)

    1      2      3      4      5      6      7      8      9     10  11
1.186  0.826  0.900 -0.608  0.297 -0.992  0.754 -0.564 -0.188  0.543 -2.575

Studentized residuals: Flag if absolute value > 2; No. Ireland Is extreme

> rstudent(BritRegions)

    1      2      3      4      5      6      7      8      9     10  11
1.218  0.810  0.889 -0.586  0.281 -0.990  0.734 -0.542 -0.178  0.520 -4.732

Cook's distance: Flag if D > 0.5: No. Ireland is very extreme!

> cooks.distance(BritRegions)

    1      2      3      4      5      6      7      8      9     10  11
0.11410  0.03652  0.04373  0.02360  0.00474  0.14733  0.04665  0.07749  0.00182  0.06892  1.74723
Plot(BritRegions) produces 4 plots, including this one

Remove Northern Ireland

```r
> NoIreland = subset(BritishRegions, subset=Region!="NorthernIreland")

> NoIre <- lm(Alcohol ~ Tobacco, data = NoIreland)

Coefficients:

            Estimate Std. Error t value  Pr(>|t|)  
(Intercept)  2.041      1.001    2.04   0.0759 .
Tobacco      1.006      0.281    3.58   0.0072 **

Residual standard error: 0.446 on 8 degrees of freedom
Multiple R-squared:  0.615,  Adjusted R-squared: 0.567
F-statistic: 12.8 on 1 and 8 DF,  p-value: 0.00723
```
Without Northern Ireland, Scotland is a problem

Reminder from Lecture 3
Reasons for outliers and what to do

- A mistake was made. If it’s obvious that a mistake was made in recording the data, or that the person obviously lied, etc., it’s okay to throw out an outlier and do the analysis without it. For example, a height of 7 inches is an obvious mistake. If you can’t go back and figure out what it should have been (70 inches? 72 inches? 67 inches?) you have no choice but to discard that case.
- The person (or unit) belongs to a different population, and should not be part of the analysis, so it’s okay to remove the point(s). Example: predicting house prices, if data set has a few mansions (5000+ square feet) but other houses are all smaller (1000 to 2500 square feet), then it makes sense to predict sales prices for the smaller houses only. In the future when equation is used, it should be used only for the range of data from which it was generated.
- Sometimes outliers are simply the result of natural variability. In that case, it is NOT okay to discard them. If you do, you will underestimate the variance.