Conditions for Inference

RECALL that the conditions for statistical inference in regression were:

– Same SD, for all levels of X
– Independent observations
– Normal distribution of Y for each fixed X
– Random sample

In addition it must be reasonable to fit a line to the data

Residual Plots

→ What are two plots that would help use check the linearity assumption?

→ Which plot is better? Why?

→ What do you plot to create a residual plot?

→ What do you want the residual plot to show?

Example: Birth Time (continued)

→ What does this residual plot suggest?
What other shapes should you look out for in a residual plot? Draw an example.

What does a funnel shaped residual plot suggest about your regression data?

Example: Airfare (continued)

What does the residual plot suggest?

Normal Plots of Residuals

What assumption does the normal plot of the residuals satisfy (easy question I am not tricking you.)

Transformations

Why do we use transformations to our data in regression?
What variables get transformed?

List some common transformations

Example: Remember the highly debated presidential race of 2000? Recall that one of the highly criticized results of this election was the voting irregularity found in Palm Beach County, Florida. Democrats argued that Pat Buchanan received far too many votes, which should have been for Al Gore, all because of the confusing butterfly ballot. The variables in the file 2000pres.mtw are the 2000 election votes for each major party candidate: Gore, Bush, Buchanan, Nader, and total votes; by county in Florida. One of the justifications for this abnormally large number of Buchanan votes (from republicans obviously) was that Palm Beach is one of the largest counties in Florida, so there should be more votes for every candidate when compared to other counties in the state. To investigate this claim we will use regression to explore the relationship, if any, between total number of votes (ie. a surrogate for county size) and votes for Buchanan.

Which observation is Palm Beach?
→ What is the point of taking Palm Beach out of the analysis and rerunning the regression?

Regression Analysis: BUCHANAN versus Total

The regression equation is
BUCHANAN = 84.1 + 0.00151 Total

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>84.09</td>
<td>17.76</td>
<td>4.73</td>
<td>0.000</td>
</tr>
<tr>
<td>Total</td>
<td>0.0015120</td>
<td>0.0001183</td>
<td>12.78</td>
<td>0.000</td>
</tr>
</tbody>
</table>

S = 119.745  R-Sq = 71.9%  R-Sq(adj) = 71.4%

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>2342846</td>
<td>2342846</td>
<td>163.39</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>64</td>
<td>917686</td>
<td>14339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>65</td>
<td>3260532</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

→ What do the regression results suggest? Does total appear to be a significant predictor of votes for Buchanan?
What does the residual plot suggest?

What does the normal plot of the residuals suggest?

What can we do to fix this problem?

Why does the log seem like an appropriate transformation in this case?
Regression Analysis: LOGBuch versus LOGTotal

The regression equation is
LOGBuch = -1.09 + 0.703 LOGTotal

Predictor    Coef    SE Coef     T      P
Constant  -1.0866   0.1640  -6.62  0.000
LOGTotal   0.70349  0.03610  19.49  0.000

S = 0.189026    R-Sq = 85.6%    R-Sq(adj) = 85.4%

Analysis of Variance

Source     DF   SS    MS   F    P
Regression  1  13.569  13.569 379.76  0.000
Residual Error 64  2.287   0.036
Total       65  15.856

→ Do the transformed regression results suggest we have a decent model?

→ What does the transformed regression residual plot suggest?
What does the transformed normal plot of the residuals suggest?

SO...Based on the voting pattern for the rest of Florida (excluding Palm Beach) in the 2000 election, what do we predict for number of votes for Buchanan, based on the total number of votes for Palm Beach.

Palm Beach actually cast a total of 432286 votes, which Log(432286) = 5.63577

So we would predict
LogBuch = -1.09 + 0.703 (5.63577) = 2.8781

Which equates to $10^{2.8781} = 755.26$ or 756 votes.

How many votes did Buchanan actually get in Palm Beach?

In fact, a 95% prediction interval suggests that we are highly confident based on the voting pattern in the rest of Florida, Buchanan should have received between $10^{2.4888} = 308.18$ to $10^{3.2673} = 1850.55$ votes.

Predicted Values for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.8781</td>
<td>0.0472</td>
<td>(2.7838, 2.9723)</td>
<td>(2.4888, 3.2673)</td>
</tr>
</tbody>
</table>