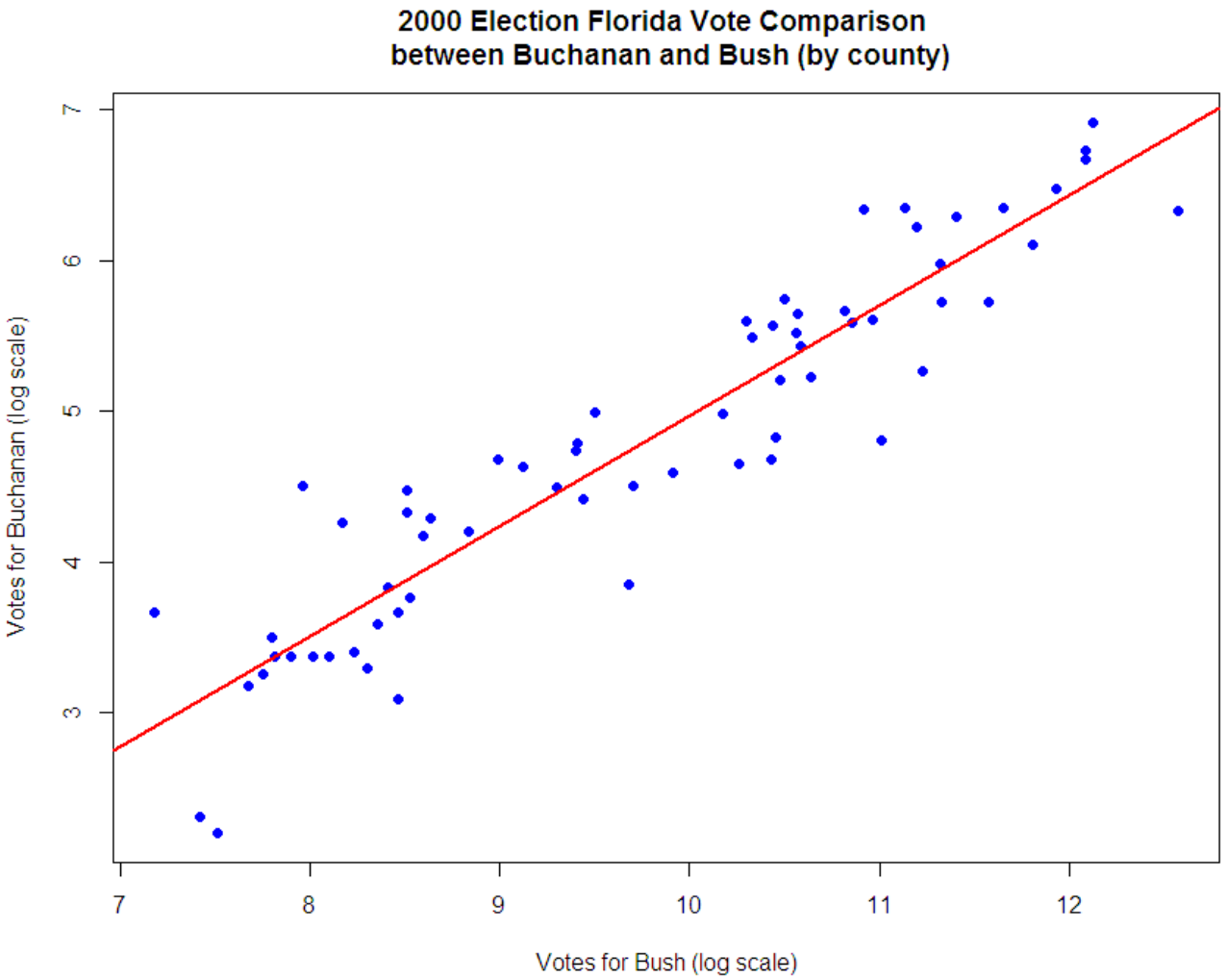
Andrew Goldstein
March 21st, 2012

Bush, Buchanan, and Palm Beach County

There are 67 counties in the full data set but – in order to explore the hypothesis that votes for Bush can be used to explain votes for Buchanan in all counties other than Palm Beach – I plotted the points and determined a linear model for the subset z of 66 counties excluding Palm Beach.

I chose to use a log scale in order to correct for the differences in residuals due to heavy tails and huge discrepancies in the size of counties; in this case, the log scale correction is important because it ensures that there will be *homoscedastic* errors.

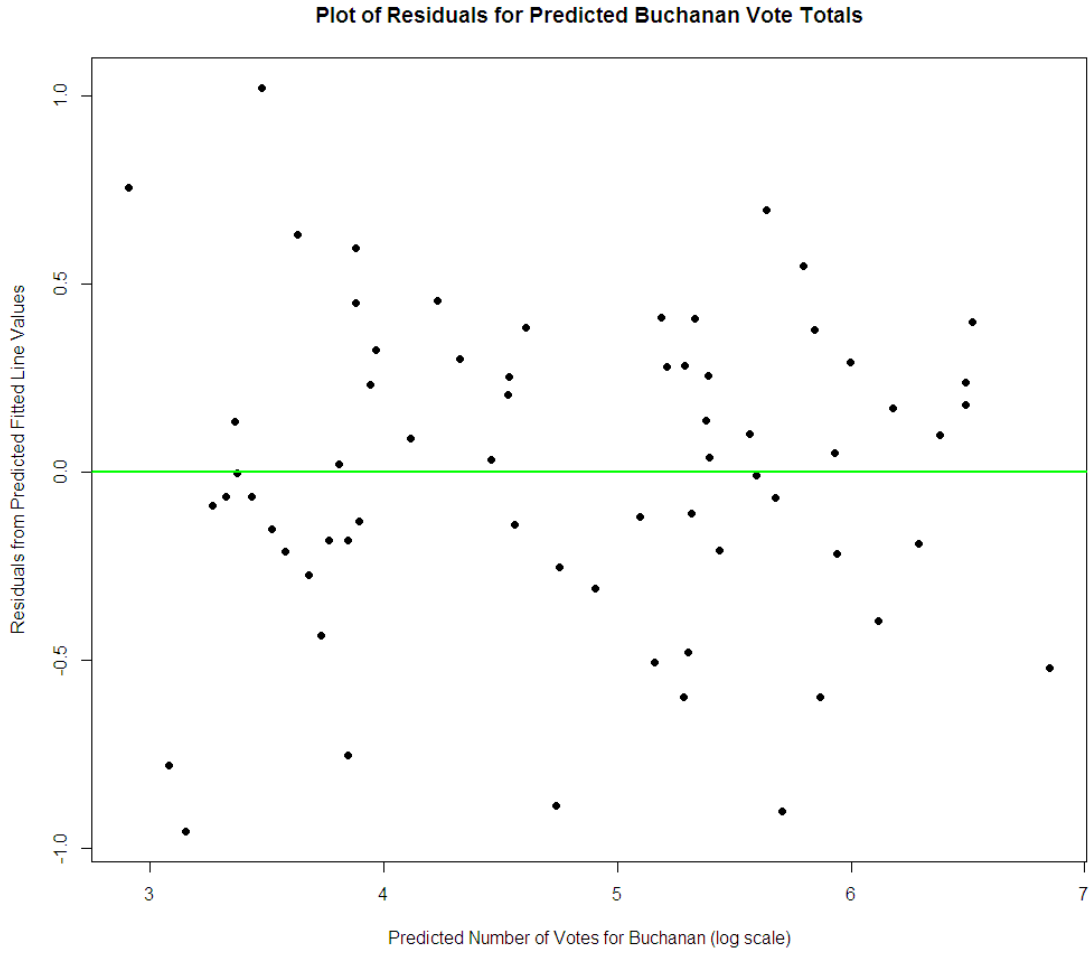
The R-squared value for this linear fit is a respectable 0.86.



Andrew Goldstein
March 21st, 2012

The figure at right shows the residuals for the fitted line from the figure on the previous page. In addition to a respectable R-squared value as established previously, the green line in the figure simply shows that the residuals have no overriding pattern and are truly random, hinting at a "good" linear fit.

We can also learn from the plot below that all counties in Florida are somewhat similar to one another in terms of their preferences for Bush and Buchanan. Of course there are large differences in pure number of votes due to population which are obscured by the log scale format, but the larger point is that there is less variation and fewer outlier counties than one would expect.



However, the expected (fitted) value for Palm Beach County does not line up with the observed data:

```
> as.numeric(predict(out, x)[x$pb])  #EXPECTED (FITTED VALUE)
[1] 6.384143
> log(x[67,2])  #OBSERVED (ACTUAL VALUE)
[1] 8.133587
```

For the case of Palm Beach, the size of the residual is the difference between the two values above, or roughly 1.749. If we compare this residual to the plot on the previous page, it quickly becomes evident that this residual is much greater than all of the others.

The question then becomes, "Could this residual have occurred by chance?" It order to tackle this question, I chose to divide the residual for Palm Beach County by the standard deviation of all the residuals in order to come up with a z-value.

```
> as.numeric(log(x[67,2]) - predict(out, x)[x$pb])/sd(out$resid)  # Z-VALUE
[1] 4.199756
```

Then, I created a test to find the maximum residual for a set of 67 randomly (normal) generated residuals and repeated the task 100000 times.

```
> test <- rep(NA, 100000)
> for (i in 1:100000) {
+   vals <- rnorm(67)
+   test[i] <- max(abs(vals)) }
```

The number of maximum residuals at least as extreme as that observed for Palm Beach County was only 154 leading to an effective p-value of 0.00154.

```
> mean(test > 4.199756)
[1] 0.00154
```

In conclusion, I will say that it is possible that this residual (and by extension, the vote tallies in Palm Beach County) could have occurred by chance, but I find that explanation to be extremely unlikely. The p-value of 0.00154 goes well beyond the threshold required to be statistically significant.