

PS 787 Assignment 1 (Sept 26, 2007)

The data in `FLdat1.csv` are for each of the 67 counties of Florida. The `.csv` file contains eight variables:

1. county name (this column lacks a name in the first, header row)
2. "Votes00": total of votes recorded for a presidential candidate in the 2000 election
3. "Reform00": votes recorded for the Reform party presidential candidate (Buchanan) in 2000
4. "Votes04": total of votes recorded for a presidential candidate in the 2004 election
5. "Reform04": votes recorded for the Reform party presidential candidate (Nader) in 2004
6. "Population": county population as of 2000 Census
7. "BlackProp": proportion of county population with race black in 2000 Census
8. "CubanProp": proportion of county population with Cuban national origin in 2000 Census

Notoriously in the 2000 election, the Reform vote in Palm Beach County was a large outlier: Buchanan received about nine times as many votes as he should have due to a defective ballot design. Each of the three demographic variables (Population, BlackProp and CubanProp) has a strongly skewed distribution across the counties.

These data may then be a good setting to explore the topic of disproportionate influence for particular observations in a linear regression analysis. Suppose we are interested in the relationship at the county level between voting for Reform in 2000 and voting for Reform in 2004. The **R** program `assign1.R` runs three regression models. In all three models the dependent variable is `log(Reform04)` (that's the natural logarithm). The regressors are respectively

1. `log(Reform00)`
2. `log(Reform00), log(Population)`
3. `log(Reform00), log(Population), CubanProp, BlackProp`

An ANOVA table appears at the end of `assign1.R`. Results from running that program via the command line (in Linux)

```
R CMD BATCH --no-save assign1.R
```

are in `assign1.Rout`.

What if any concerns might we reasonably have about the analysis due to one or more observations being exceptionally influential? Does an influence analysis depend on the specified set of regressors? What if anything might one do to use a linear regression analysis to get a believable answer to the question about the county-level relationship between presidential Reform votes across the two elections? For all these questions, present a detailed response supported by the data in `FLdat1.csv` (you need not use **R** to do your analysis).

For the sake of this exercise, please ignore possible concerns about heteroscedasticity or serial dependence.