

PS 787 Assignment 2 (Oct 19, 2007, due Oct 31)

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

This exercise has two main aspects: count data models and specification concerns. The basic count data model is based on the Poisson distribution. Is a Poisson regression model correctly specified for these data?

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 `assign2.R` runs four regression-like models: log-linear OLS regression; Poisson regression; negative binomial regression; and an overdispersed log-linear generalized linear model. In all the models the dependent variable is `Reform04` and the regressors are `log(Reform00)` and `log(Population)`. Results from running that program via the command line (in Linux)

```
R CMD BATCH --no-save assign2.R
```

are in `assign2.Rout`.

The `assign2.R` program uses the iteratively reweighted least squares algorithm associated with generalized linear models to estimate the models (**R** functions `glm` and `glm.nb`). The last part of the `assign2.R` program has code that implements a few of the overdispersion tests.

Is the Poisson model correctly specified? If not (or if so), how can one obtain consistent estimates for the mean and covariance matrix of the model's parameter estimates? If the Poisson model is abandoned, is a negative binomial regression model correctly specified?

Extra credit: are the specification results linked to the influence of different observations? Does adding more covariates (specifically `CubanProp` and `BlackProp`) change things?

For all these questions, present a detailed response supported by the data in `FLdat1.csv` (you need not use **R** to do your analysis).

addition 1 (Oct 20): The `assign2a.R` file contains everything in `assign2a.R` plus commands to use the `nlm` function to estimate the Poisson regression model directly by maximum likelihood. The `nlm` function minimizes, so the approach in `assign2a.R` is to give `nlm` a function that sums the negative of the Poisson log likelihood.

`assign2a.R` also computes the simple difference between elements of the observed and expected information matrices, summing the lower triangle over observations. This is not the test statistic specified in White's test, but it can be used to compute that statistic. The details of doing that are left as part of the exercise.