

Week 7: Negative Binomial and Stuff

Exercise 1: Swedish traffic accidents

In 1961 and 1962, Sweden conducted an experiment to evaluate the effectiveness of speed limits on reducing traffic accidents. For the same 92 days in each year, some days were subject to speed limits while others were not (speed limits were not the norm on Swedish motor ways at the time). The resulting data (`library(MASS); data(Traffic)`) are counts of accidents on each day.

- (a) Model the effect of the speed limit on accidents, using a Poisson regression. For this exercise, you can ignore the other prediction variables.
- (b) Model the effect of the speed limit, using a negative binomial regression. Compare the negative binomial model to the Poisson model.
- (c) Plot your predictions, from both models, and compare them to the actual data. Use the same method of comparison that I used in the lecture, for the salamander data. The code to produce that comparison is in the `week7.r` file.

Exercise 2: UC Berkeley admissions and gender bias

The data contained in the file `ucb.csv` are counts of male and female PhD applicants admitted to various departments at UC Berkeley in the 1970's (Bickel, P. J., Hammel, E. A., and O'Connell, J. W. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187:398–403). This is a rather famous data set, featuring in debates about the evidence of gender-bias in graduate admissions. It also provides an example of something known as “Simpson's paradox.”

- (a) Model the probability of being admitted, as a function of only gender, across all departments. That is, the only prediction variable to use is the gender of the applicant. What does your model suggest? Is there evidence of gender-bias? What is the change in odds of being admitted, if one is female rather than male?
- (b) Now model the probability of being accepted, using gender as well as dummy variables for the various departments. The departments have been given aliases “A” through “F”, but these are real departments, each making autonomous admissions decisions. Is this model better than the model in (a)? What does your new model say about the question of gender-bias in admissions? What is the new change in odds of being admitted, if one is female? Can you explain the discrepancy between your results in (a) and your results here?

Exercise 3: Oceanic technology

The data in `oceanictools.csv` are 10 historic Oceanic societies, their historic populations (`Population`), contact rates with other islands (`Contact`), and most importantly their total tool counts (`Total.Tools`). A much-debated hypothesis in evolutionary anthropology has it that the size and complexity of a society's toolkit is positively correlated with population size—because larger populations are both subject to less drift that might remove specialist knowledge, as well as have more learners, increasing the chance that at least one of them successfully learns a complex tool. (See: Henrich, J.

2004. *American Antiquity* 69:197-214.) A corollary hypothesis has it that small populations in frequent contact with other groups will be buffered against technology loss, because contact increases effective population size. You will use these data from Oceania to evaluate these two hypotheses.

(a) Plot the `Total.Tools` against `Population`. What shape is the relationship? Construct a new data column that contains the logs of `Population`. Plot this against `Total.Tools` now. What shape is the relationship? Note that you can use:

```
identify( log(d$Population) , d$Total.Tools , labels=d$Culture )
```

To label the points with the culture names, by clicking on points in turn. (Change `log(d$Population)` to just `d$Population` accordingly.) When you don't want to label anymore points, press ESC on your keyboard.

(b) Model the counts in `Total.Tools`. Decide which probability distributions are plausible and construct models that may include either or both population and contact rate. These models may be linear or non-linear, with respect to how the mean of `Total.Tools` relates to `Population` or `log(Population)`. Note that there is very little data (10 rows!), so go easy on parameters. Compare these models, using both AIC and the more-conservative AICc. Interpret the results.

(c) Plot your model-based predictions. What accounts for the better fit of the best model(s)?