Chapter 14: Count Models

Exercises

Brian Fogarty

Exercise 1

a. Using the `citations_twitter.csv` dataset, create a smoothed density plot of `citations` for only tweeted articles. Does the plot suggest any potential problems that we should make sure to test?

b. Run a Poisson regression model (PRM) with `citations` as the outcome variable and `womanauthor`, `fullprof`, and `retweets` as predictors. Briefly discuss (one sentence) what predictors are statistically significant.

c. Test for overdispersion in the PRM and discuss the results.

d. Run a negative binomial regression model (NBRM) with `citations` as the outcome variable and `womanauthor`, `fullprof`, and `retweets` as predictors. Briefly discuss (one sentence) what predictors are statistically significant. What differences exist, if any, between the PRM and NBRM for statistically significant predictors?

e. Compare model fit statistics for the PRM and NBRM. This includes performing a likelihood-ratio test and comparing log-likelihood and AIC values. Based on the model fit statistics and the overdispersion test (from 'c.'), which model is preferred?

f. Using the preferred model (from 'e.'), calculate and interpret predicted counts (i.e., predicted article citations) for the lowest and highest values of all statistically significant predictors. (If there are no statistically significant predictors, skip this exercise.)

g. Using the preferred model (from 'e.'), create a predicted probability plot for each statistically significant predictor with the `ggpredict()` and `ggplot()` functions . If plotting `womanauthor` and/or `fullprof`, use the `geom_pointrange()` function to plot the predicted count estimate with confidence intervals (as in the Chapter 14). If plotting `retweets`, use the `geom_smooth()` function to create a curvilinear line (as in Chapter 13 and the online supplemental materials for Chapter 14). Briefly discuss the plot(s) (2-3 sentences). (If there are no statistically significant predictors, skip this exercise.)

Exercise 2

This exercise uses a dataset on the number of news articles in international newspapers on the 2014 Ferguson, Missouri protests following the police killing of Michael Brown. The dataset (`Ferguson International News.csv`) includes the number of articles on Ferguson (`amount`) in the major newspaper in 57 countries and the countries' Gini coefficient (`gini`), level of ethnic fractionalisation (`efrac`), and a measure of societal safety (`soc_safe`). The Gini coefficient measures a country's level of income or wealth inequality, where higher numbers indicate greater inequality.[1] The level of ethnic fractionalisation measures ethnic heterogeneity in a country, where higher numbers indicate greater ethnic heterogeneity.[2] Societal safety is an index that

[1] Data from the World Bank (2014).
[2] Data from Fearon and Laitin (2003).

measures the safety and security of a country, where higher numbers indicate lower levels of safety and security.[3]

a. Read-in the `Ferguson International News.csv` dataset and examine the mean and variance of the variable `amount`. What does the mean and variance tell you about potential overdispersion?

b. Create a smoothed density plot of `amount`. Does the plot suggest any potential problems that we should make sure to test?

c. Run a Poisson regression model (PRM) with `amount` as the outcome variable and `gini`, `efrac`, and `soc_safe` as predictors. Briefly discuss (one sentence) what predictors are statistically significant?

d. Test for overdispersion in the PRM and discuss the results.

e. Run a negative binomial regression model (NBRM) with `amount` as the outcome variable and `gini`, `efrac`, and `soc_safe` as predictors. Briefly discuss (one sentence) what predictors are statistically significant. What differences exist, if any, between the PRM and NBRM for statistically significant predictors.

f. Compare model fit statistics for the PRM and NBRM. This includes performing a likelihood-ratio test and comparing log-likelihood and AIC values. Based on the model fit statistics and the overdispersion test (from 'd.'), which model is preferred?

g. Using the preferred model (from 'f.'), calculate and interpret predicted counts (i.e., predicted articles) for the lowest and highest values of all statistically significant predictors. (If there are no statistically significant predictors, skip this exercise.)

h. Using the preferred model (from 'f.'), create a predicted probability plot for each statistically significant predictor with the `ggpredict()` and `ggplot()` functions . Briefly discuss the plot(s) (1-2 sentences). (If there are no statistically significant predictors, skip this exercise.)

For the answers see **Chapter 14 - Answers to Exercises**.

[3]Data from Global Peace Index (2015).