Week 8: discrete data and GLM(M) regression

The Assignment

20 March 2019

Data

Background

This assignment uses a dataset from the Cleveland Police Department that contains several years' worth of all charges filed within 800m of the two major league sports arenas in the city (Fig. 1). We were interested in how athletic events affect crime rates in the neighborhoods of these venues. Thus, the dataset includes whether more charges were filed on a game day or not, the type and time of day for each charge, and for the venue with more than one event[1], it includes what sport was being played on (many of) the game days.

Structure

Load the Cleveland crime data and report its structure.

Compare discrete data with simple variance structure

For this section, use only data from Gateway/Progressive Field.

Game days vs. non-game days

Produce a bar graph showing whether there are more charges on game days than non-game days.

Crime by event type

According to these data, which sports are played in the Gateway/Progressive Field complex? Which generates the greatest amount of charges? Produce a bar graph comparing the number of charges per event type, test the difference statistically, and interpret the result. *Any ideas why this pattern might occur?*

Compare discrete data with non-independent variance

What special considerations must we give were we to model data from both venues together?

Data presentation

- Present a table of the number of charges on game days and non-game days per venue
- Present these same data in a bar graph.

[1] The Gateway Complex includes both Progressive Field, which hosts Cleveland's MLB baseball team, and the Quicken Loans Arena hosts the Cavaliers NBA team, while FirstEnergy Stadium is just for Browns football.

Statistical testing and interpretation

Fit an appropriate model, or models, to test the hypothesis that more charges are filed on game days than non-game days in the neighborhoods of both major league athletic venues in Cleveland. Note that we want to make a general statement about the effect of game day on crime, not statements unique to each venue.

Respond to the following:

- In general, describe two important considerations you must give to ensure you fit the appropriate model that relate to the nature and necessary error structure of the data.
- Specific to these data, what type of regression model should you fit? What is the best distribution?

Going further

Does the number of charges filed vary with charge type? Present graphical and statistical evidence to support your claim.

- What alterations, if any, might you make to your modeling strategy?

Now focus on just the three most-reported charge types:

- Are there significant differences between the number of charges across charge types?
- Is game day a meaningful variable in the relationship between the number of charges and charge type? Provide statistical evidence without a P-value.
- Generalize the effect of each of the three charge types. Which is most likely to occur on game days? Present evidence without a P-value.

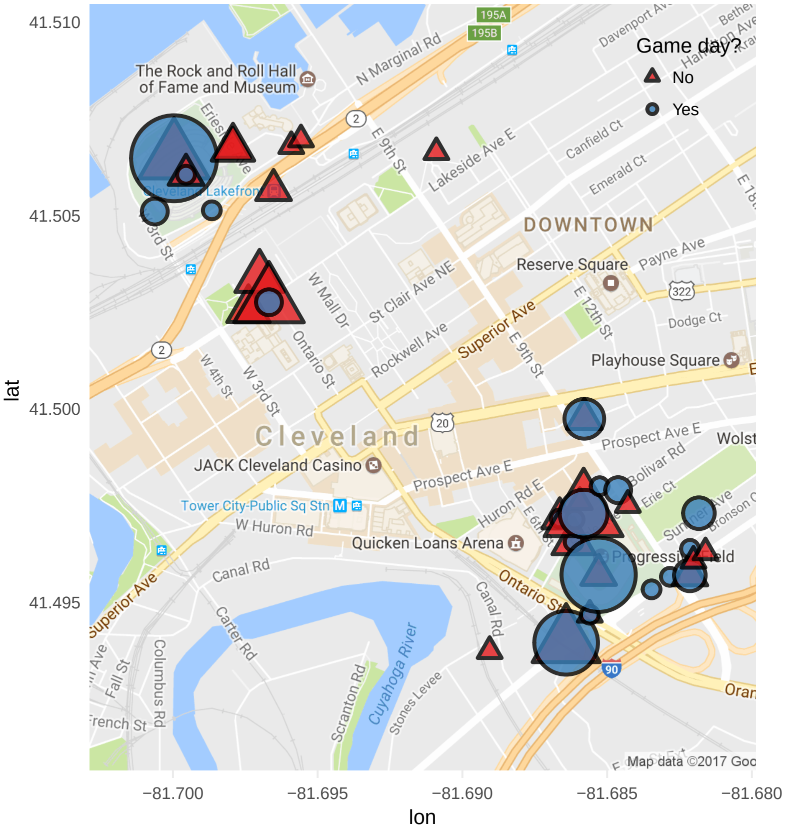

Figure 1: A map of downtown Cleveland showing crime clusters in the neighborhood of each major league sports venue. Blue circles denote charges filed on game days while red triangles denote charges filed on non-game days. Symbol size scales with number of charges reported at each location.