Applying GAM To the Homeless Data

Spring, 2017

1 Background

You have seen and analyzed these data for the practice assignment. The goal of this assignment is to have you apply GAM. The first few sections of the report should be a slam dunk. We went over that material in class.

As before, there should be the following sections in your research report:

1. Introduction: Statement of The Problem
2. Description of The Data (including how the data were generated)
3. Cleaning Up The Data (e.g., dealing with unreasonable values)
4. Analysis of Univariate Statistics (statistical and substantive issues)
5. Analysis of Bivariate Statistics (statistical and substantive issues)
6. Multivariate Analysis (statistical and substantive issues)
7. Summary and Conclusions

Also as before, think about the features of the tools you are using and justify your decisions. On the substantive issues, what are you learning with respect to your problem statement? Be your own harshest critic and respond. Why should anyone accept what you did with the data and believe your empirical results?

2 Homelessness in Los Angeles

Here's a recap. Public policy for the homeless requires information on where they are and how numerous they are. The information is needed to determine what services are needed where. These data are from a random sample of 509 census tracts out of over 2000 in Los Angeles. The response variable is the homeless street count for each tract undertaken on two successive

days/evenings. For this assignment, your job is to consider what features of census tracts are related to homelessness.

The variables are as follows.

1. StreetTotal – The number of homeless individuals in a census tract
2. MedianIncome – Median household income in a census tract
3. PropVacant – The proportion of vacant residential dwellings in a census tract
4. PropMinority – The proportion of minority residents in a census tract
5. PctCommercial – The percentage of a census tract zoned for commercial enterprises
6. PctIndustrial – The percentage of a census tract zoned for industrial enterprises

3 Issues to Consider

Perhaps most important, using the "wrong-model" approach, you need to decide whether you are doing a Level I, Level II, or Level III analysis. Then you need to explain and justify your decision(s). What assumptions, if any, are you making? If you are making assumptions, how credible are they?

When you interpret you statistical results, don't just say in words what anyone could read off of your output. Say in words what the results mean for the substantive question(s) being addressed. Be precise. If there is an association to report for parametric association, for example, don't just tell me the sign. Tell me what coefficient you are using and exactly what it conveys about the relationship. When you are using visualizations, explain what you see and how that translates into an interpretation.

BIG Hints – If you decide to go with a Level II analysis, exploit sample splitting. Arrive at the model or models you like working between the training data and the evaluation data. Then move that model(s) to the test data with the chosen values of your tuning parameter fixed. The tests reported in the conventional printout from the test data are not quite correct under the wrong model perspective, but until we do the bootstrap, they are the best you have and should be reasonably close to correct. I recommend you use *gam* from the *gam* library. But *gam* from the *mgcv* library is fine too (but trickier to use correctly).

Around 5 single-spaced pages of text should be adequate. Tables, figures and R-code would be extra. Number your pages, figures and tables. Refer to table numbers and figure numbers in the text. There are lots of examples of the kinds of discussion necessary in Chapter 2. Likewise, you will find in chapter 2 templates for the code you need.