Midterm Exam I

Advanced Methods for Data Analysis (36-402)

Due Thursday March 7, 2019, at 6 PM

Remember to read the guidelines regarding the preparation and submission of a data analysis report. In particular, you are **not** allowed to collaborate with any other students regarding this work.

Do not forget the highlighting and numbering step. If you do not know what I am talking about, then read the guidelines.

Your report should be **at most** 15 pages. Be sure to leave enough time for writing. We expect to see a clearly written report.

Your data exam (**report and code**) must be submitted through Canvas by 6:00 PM on Thursday, March 7. **There will not be any extensions.**

Important announcements and clarifications regarding this exam will be made via Canvas. Make sure you are reading your Andrew email on a regular basis.

We will have regular lectures on Tuesday March 5 and Thursday March 7 on HW 7 material.

The Data

The City of Chicago maintains a database of all criminal reports filed by their police department, all of which since 2001 is open to the public for their inspection. These include cases in which no arrest was made, and extend from the lightest criminal charges up to and including homicide. A great deal of public policy in recent years has focused on using these reports in "predictive policing", or forecasting criminal activity in a way that allows for a better deployment of police resources. We have collected narcotic-related crime reports across 2012 per Census block group. Within each block group, we have variables from the 2010 US Census and the 2011 American Community Survey that provide additional information. The counts of narcotic-related crimes are split in two: cannabis and non-cannabis related cases.

Within the file DA-exam-data.RData is a data frame called *chicago*, with 2102 rows - representing each block group with at least one report - and 13 columns. Information on each column follows.

- *poptotal*: the total count of members of the block group.
- *pctWhite, pctBlack, pctAsian*: the proportion members of these ethnic groups within the block group.
- *income.male, income.female*: the average income earned by members of that gender in the block group.
- *age.male, age.female*: the average age of members of that gender in the block group.
- *longitude, latitude*: the (x,y) coordinates of the block group.
- *Ward*: division of the city of Chicago based on political and geographic delineation.
- *CrimeNC*: the total number of noncannabis related crimes in Chicago during 2012.

- *CrimeC*: the total number of cannabis related crimes in Chicago during 2012.
- *Zone*: 1 if block is north to the river, 0 if block is south to the river (variable added by the 402 Team).

The Goals

Your goal is to understand how demographic and geographic factors relate to narcotic-related crime in Chicago. **(1)** Does narcotic-related crimes depend on demographic and geographic factors? **(2)** Moreover, are cannabis and noncannabis-related crimes correlated?

Your task is to build stable, interpretable, theoretically valid regression models to predict narcotic crime in chicago that address these research questions.

Specifically, you must address the following issues in your report along with all other issues necessary for describing and justifying a statistically valid analysis:

Introduction

Clearly state the research questions and objectives of your study.

Exploratory Data Analysis

Part 1

Create a new variable in the data set called *'CrimeTotal'*, the total count of narcotic-related crimes per block. Describe all variables in the data set through univariate EDA. Summarize them (numerically/graphically as appropriate). How many observations do you have? Are there any interesting patterns? Are there any outliers?

Part 2

Do multivariate EDA (e.g. scatter and side-by-side box plots as appropriate). Describe any trends or interesting features that you see. Remember to analyse the relationship between all available variables and the three responses (*TotalCrime, CrimeNC* and *CrimeC*).

Part 3

For each type of crime, identify the top 5% highest, and the bottom 5% lowest crime rate blocks. Plot them using longitude and latitude in the (x,y) axes. Do you see a geographical pattern for high crime rates? And for low? Are the patterns different for the two types of narcotic-related crime?

Initial Modeling & Diagnostics

Part 4

Using insights you gained from the EDA, construct two candidate multiple linear regression models for the total crime count (cannabis and non-cannabis crimes added together) per block. Explain your choices for how you code each variable; including potential transformations and whether you decide to treat discrete variables as continuous or categorical. Explain your choices for which variables to include in the models.

Part 5

Do model selection using a statistically rigorous criterion to minimize prediction error, and state the model chosen. Attach measure of uncertainty to your estimated prediction errors and interpret these as appropriate.

Part 6

Present model diagnostics. Discuss possible improvements and modifications to your model to address any violations of the model assumptions.

Part 7

Are any transformations needed in order to meet the model assumptions?

Part 8

Were you able to address all concerns about the model assumptions?

Results

Part 9

Regardless of the models you have constructed so far, fit a linear model with all covariates (and any transformations you believe it is appropriate). Does there seem to be a relationship between total crime rate and geographic and demographic variables? Make sure you clearly state your null and alternate hypotheses, your test statistic, and how you perform your test. *Hint:* remember that a series of single hypothesis tests is generally not the correct way of tackling this problem.

Part 10

Does being above or below the river affect the cannabis-related versus non-cannabis-related crime counts in a block tract differently? Create one linear regression model per type of crime using the `zone` variable and other covariates you deem necessary, if any. Compare the regression coefficient for `zone` for both models. State the statistical hypothesis, provide a p-value and draw a conclusion. *Hint:* there is more than one way to approach this problem, whatever strategy you take, make sure you explain your reasoning!

Part 11

Create a data set with total crime per ward and total population per ward as variables. Which are the 5 wards with the highest count of crime? Using this data set, fit a model with count of crime per ward as the response and total population as a covariate. Rank the wards according to the residuals. What does that tell you about the wards with highest crime rates? Is correcting by population size reasonable? Why? *Hint:* Use the group_by and summarise functions from the dplyr package to create the data set.

Part 12

Is there a relationship between cannabis and non-cannabis related police reports in each block group? To answer this question, construct a model with one type of crime as the response and the other as the predictor. What happens to this relationship when you control for the other variables? What does this mean? *Hint:* you can use transformations where appropriate.

Conclusions/Discussion:

Part 13

Discuss your results with respect to the research hypotheses. Summarize your main findings in the analysis. Discuss possible reasons for these findings. Make some recommendations for future work or studies but be brief.