The University of Texas at Austin
Department of Electrical and Computer Engineering

MIS382N: Business Data Science — Fall 2019

Problem Set Five

Caramanis/Dimakis Due: Thursday October 17th, 3:00pm 2019.

Problem 1

This problem walks us through a problem discussed in detail in the Multi-level regression book written by A. Gelman and J. Hill. NYC has a program known as stop-and-frisk. Relying on a 60's era ruling, the law allows an officer to search someone without arrest, and without probable cause, if the officer believes s/he might be in danger because of a hidden weapon. Much has been written about this, as it has come under significant scrutiny for being discriminative and allowing (even encouraging) racial profiling. You can read a summary of it at this Wikipedia page: `https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City`.

The data you will download contain information about the number of traffic stops, reported per precinct (75 total precincts), along with the ethnicity as reported by the police officer. These data have kept only three ethnicities: white, black and hispanic. The data set also has data on arrest rates in the previous year, broken down by four types of crimes; and the total population levels per precinct per ethnic group.

1. Download and load the data in the file `NYC_stop_and_frisk.dat`, uploaded to Canvas.
2. What fraction of the total stops correspond to "white/back/hispanic"? What fraction of the population corresponds to "white/black/hispanic"?
3. Use a Poisson regression to model the number of stops, controlling for ethnicity and using the number of past arrests as an exposure input.[1]
4. According to the output of your model, what fraction fewer or more stops does each ethnicity have with respect to the others, in proportion to arrest rates of the previous year? Note that you can just pick a baseline ethnicity and just compare everything to that.
5. Next, add the 75 precincts, and again solve the Poisson regression model.
6. Now, controlling for precincts, according to your model, what fraction fewer or more stops does each ethnicity have with respect to the others, in proportion to arrest rates of the previous year? (Again, just report with respect to a chosen ethnicity as a baseline).

Problem 2

In this problem you will play with the idea of compound models. I have created a data set (entirely fake!)[2] of 2012 salaries in the NBA, of 10,000 basketball players that were in high-school

[1] Recall that you use an exposure input as follows: Let u_i be the exposure you are using for data point i – in this case, u_i is the number of arrests in the previous year; in the traffic accident example, one choice of exposure is the number of cars that pass through intersection i. Then you are modeling: $y_i \sim \text{Poisson}(u_i\theta_i)$ where $\theta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

[2] So fake, that there are more high school basketball players in 2011 that made it to the NBA in 2012 according to this dataset, than there actually were total NBA players in 2012...

in 2011: `nba_cc_fake_data.csv`. Note that the vast majority of the salaries are equal to 0 because the vast majority of these high-school players did not make it to the NBA and hence their NBA salary equals zero.

There are three features you will use: height (in inches), average points scored during the last year in high school competition, and a scoring from 1-10 of the competitiveness of the league these players played in, with 10 being the most competitive.

The goal is to build a model to predict the NBA salary of a high school baller.

- Explain why linear regression is not appropriate, given the nature of the data.
- Try least squares regression, anyway. How well do you do?
- You will next build a *composite* model. You will first predict the probability that a player actually makes it to the NBA at all, and then you will build a model to predict the salary of a player, conditioned on the fact of making it to the NBA.
  - Build a model that predicts the probability of making it to the NBA.
  - Do a train-test split of 8000/2000 points, train your best model on the training set, and compute the AUC on the test set.
  - Now, build a model to predict the salary. Note that you may wish to consider a non-linear transformation of your data. What is your R^2 score on the test set?
- Compute the expected NBA salary of a high school basketball player who is 6' 6" tall, is averaging 46 points per game, and is playing in the second most competitive league (comp = 9), according to your model.

Problem 3 (Optional)

In class we talked about the Poisson regression problem. Recall that this is appropriate for count data, and the model is:

$$\begin{aligned} y_i &\sim \text{Poisson}(\theta_i) \\ \theta_i &= \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}). \end{aligned}$$

The parameters $\boldsymbol{\beta}$ are chosen using the principle of max likelihood.

Explicitly write down the log likelihood, and show that the max likelihood value of $\boldsymbol{\beta}$ is the one that solves:

$$\min_{\boldsymbol{\beta}} : \sum_{i=1}^{n} \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - y_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Problem 4 (Optional)

Show that this function is *convex* as a function of $\boldsymbol{\beta}$.