Case Study 1 Writeup

Introduction

The goal of this project is to investigate the heterogeneity of price per milligram of morphine by location and its relationship with other factors in a data set collected from `streetrx.com`. Made online since 2010, StreetRx has been collecting citizen-reported street price data of diverted pharmaceutical substances, such as morphine. Since the price of morphine provides information about the demand, availability, and potential drug abuse, the investigation of this project could reveal insights of the morphine price over various locations and other indicators that are helpful in health surveillance for drug abuse and other health-related issues. To this end, we have built a hierarchical model on morphine price with location grouping variable through a the data analysis and modeling process performed in the following orders. First, since the data set entries had inconsistent naming problems in the location information and other issues, data cleaning was conducted to generate a data set ready for modeling. Second, we performed exploratory data analysis to understand the distributions of variables and their potential relationships. Third, we looked into the heterogeneity of the response variable by different scales of locations and chose the proper location variable as the grouping variable with random effect. Then, the other predictors, either factors or numeric variables, were selected to the model by the logical reasoning and model performance once that predictor was added. Possible interactions among predictors, which could be either fixed effects or random slopes, were also investigated in this step. Finally, we built the final model using the results from the previous sections. We checked the validness of the model assumptions and interpreted the model results, and drew the conclusion. **(Add Conclusion Here!)**

In summary, we performed data cleaning, exploratory data analysis, group variable selection, predictor variable selection and modeling, and finally model checking and interpretation to obtain the conclusion of the project. **(or Add Conclusion Here!)**

Data Cleaning

The original data set contains observations of transactions of various types of drugs. By inspecting the data set values and its structures for each predictor, we found the severla issues with noisy and corrupted data. Here is a summary of the issues and solutions to clean the data set.

First, in the original data set, `api_temp` variable specifies the active ingredient of the drug for each observation, and we need to filter the entire data set to obtain the morphine transaction observations according to the drug type specified by this variable. However, there were ambiguous labels of the **drug type** in the entries of `api_temp` variable, For example, there were categories specified as `morphine/oxycodone` rather than `morphine` only. Since we did not want observations that were related to other drug types to interfere the morphine price modeling process, we only kept the entries that exclusively listed `morphine`. This filtering process results in a new data set with 9268 records.

Second, we reformatted the **date data** in the data set. There were two fields representing date information: `yq_pdate` and `price_date`. The first field coded the year and quarter as a pseudo-continuous range. However, it was not suitable for assessing the data on a true continuous scale because it portrayed each quarter as being closer to adjacent quarters in the same year than adjacent quarters in neighboring years (i.e., fourth quarter 2019 was closer to third quarter 2019 than it was to first quarter 2020 in this coding scheme). Therefore, intead of using `yq_pdate`, we used the `price_date` to construct a true continuous scale for measuring the date of purchase. Using R's string parsing methods we parsed the date into its individual *month*, *date*, and *year*, and then creating a new field for each record that counts the *number of elapsed days since an arbitrary reference date*. In this case, we set *January 1, 2010* as the starting date, since the Streetrx data collection

began in 2010. In addition, upon inspecting the date data we observed that 3 entries were from the 1960s and an additional 11 entries were from the 2000s, which were years before the data collection was active. In this case, there was legitimate concern that the entered data might not be accurate if it was supplied long after the drug purchase event. Since a relatively small number of data were affected, we filtered out the 14 entries with dates prior to 2010, leaving us with 9254 observations.

Third, we have assessed data on the city in which the drugs were purchased, and the data set contained **city aliases and ambiguous references to some cities** other than the formal names of cities. We saw that there were 1690 unique "city" values among the 9254 observations. However, when we inspected visually we could see almost immediately that there were numerous entries with **different listed names that clearly refer to the same city**. This was because users might have used different conventions when supplying city names. For example, users might have listed either "Fort Lauderdale" or "Ft Lauderdale" to refer to the same city, resulting in two different values. We saw a range of other common data discrepancies, such as using city nicknames (e.g., "Philly" for Philadelphia), airport codes (e.g., "ATL" for Atlanta), or using the abbreviation used by major sports franchises within the city (e.g., "JAX" for Jacksonville). Various other issues were observed frequently, such as users providing redundant state information (e.g., "Des Moines, IA" instead of just Des Moines), rerring to a city by the specific neighborhood or borrough (e.g., "Brooklyn" instead of "New York") or including single character typos (e.g., "Holywood" instead "Hollywood"). We employed the following two approaches to address this issue.

In most cases, the original city could be unambiguously identified and corrected manually. First, we imported US census data which defined the official city abbreviation used by each city and cross referenced it with the listed data. Fortunately, this process covered the majority of the entries, but there were still about 300 city values supplied that were not on the list. Therefore, we manually created a new dictionary mapping the noisy values to the corrected values, in the cases identified above where the data could be definitively determined. Then, we applied this mapping to the original data to correct those entries which were non-compliant. During this process we also identified legitimate names of unincorporated areas and townships which were excluded from the original census data, so that we would not be unnecessarily throwing out data from real cities.

In other cases, the original city could not be unambiguously identified from the data given. For example, some users listed their zip code, which often crossed city lines. Others listed their county, which included multiple cities, and others listed the general metropolitan areas (e.g., "Lehigh Valley" or "Dallas - Fort Worth"). We did not want to guess, so in these cases, we replaced any remaining city names with **Other/Unknown**.

One pitfall that we had to avoid was inappropriately aggregating city data that were not related. For example, the data set contained both "Hollywood, FL" and "Hollywood, CA". If we ultimately built a hierarchical model with both state and city grouping variables, we did not want to mistakenly label data from those two places as being from different states but the same city. Therefore, we augmented our coding of the city name by appending the state as well so that each city was uniquely encoded, even if it shared its name with another city in a different state.

Then, we inspected the source data. Most entries were labeled with "Personal" or "Heard it" as the source, but there were still over 50 unique entries, too many to do serious grouping on. However, when we inspected each unique entry, we saw some common themes. First, many users entered the specific URL for various web pages they searched, and several web pages were represented repeatedly with different URLs. But more importantly, there were clear patterns in the types of sources listed. In particular, we observed that each of the sources was one of the following: (A) personal, (B) word of mouth, (C) a web forum, (D) an online black market, (E) a legal online market, (F) a web search, or (G) other/unknown. In fact, we could very easily convert the raw data to these categories by using sub-string search to find particular keywords, such as "silkroad", "bluelight", "reddit", "opiophile", "forum", "pharmacy" and various search engine names. As a result, we bundled each source into one of those seven categories, making for much simpler and more informative grouping. This data cleaning step did not change the size of the data set.

Our final data cleaning step included many minor issue corrections and basic cleaning of other predictor. First, we noticed that the state field included both the `state` values "USA" and "". We converted these values to"Other/Unknown" category since they did not provide any meaningful information. Second, we saw that `bulk_purchase` was coded as a string, which wasn't helpful, so we converted it to a numeric boolean

(0 = False, 1 = True). Third, we observed that there was nothing listed under the `Primay_Reason` field when the user did not enter the reason for the purchase. We marked such entries with missing values as "Other/Unknown" as we have with the other fields.

In summary, after the data cleaning process, we obtained a data set with 8712 observations and nine predictors, which were `ppm` (response variable), `state`, `city`, `region`, `primary reason`, `source`, `days ellapsed`, and `dosage strength`.

- `ppm`: price of morphine per milligram, numeric value.
- `city`: cities, where the transaction took place, 1654 unique factors.
- 'state": states, where the transaction took place, 56 unique factors.
- 'region": regions, where the transaction took place, 5 unique factors.
- `primary reason`: reasons for purchasing morphine, 11 unique factors.
- `source`: sources of such transactions, 8 unique values.
- `days ellapsed`: number of days passed since 01/01/2010, when the data collection process started, discrete numerical value.
- `dosage strength`: dosage strength of the purchased morphine, discrete numeric value, ranging from 1 to 200.
- `bulk`: indicator for purchased 10+ units, factor, 0 and 1.

The data analysis and modeling were performed upon the clean data set with the above predictors.

Model design

Group Variable Selection

Before determining other predictors to be included in the model, we would like to select the best location group variable which explained the heterogeneity of morphine price over that location level first. To this end, we used two approaches to achieve this end. First, by inspecting the EDA figures, (**Figures Needed Here**), we could observe that the `log(ppm)` did not differ too much across `region` variable, but there were some variations in `log(ppm)` across the sampled cities and sampled states. Therefore, we suspected that `region` might not be a good predictor explaining the heterogeneity across `region` locations. As for `city` variable, the sample sizes within each unique `city` were typically very small, so it was probably not optimal to be included as the only location group variable in the model. The `state` variable seemed to be optimal for location group variables. In addition, we also would like consider more than one location group variables in the final model.

With these suspicions and ideas in mind, we built hierarchical models with random effects on every possible combination of the three location variables, which were `city`, `region`, and `state`. There were seven models in total, which were composed by 3 hierarchical models including each location variable only, three hierarchical models each including a unique pair of the three location variables, and finally the last hierarchical model including all three location variables. Then, we compared these seven models using BIC values as criterion based on the principle that, in each comparison, the two models differed in at most one variable. For example, we could compare the BIC score of the model of random effects in `city` and `region` with the BIC score of the model of random effect in `city`. But we could not compare the model of random effects in `city` and `region` with the model of random effects in `city` and `state`. This principle could also be observed in the model comparison results, which were shown below and ranked by BIC values.

1. (`state`) < [(`state`, `city`), (`state`, `region`)] < (`state`, `city`, `region`) < (`city`, `region`).
2. (`state`, `city`) < (`city`).
3. (`state`, `region`) < (`region`).

Here are some explanations of the notations. (`state`) indicates the model with random effect in `state` only, without any other predictors. (`state`, `region`) indicates the model with random effects in both `state` and

region, no any other predictors involved. And if model1 < model2, it means that the BIC score of model 1 is smaller than the BIC score of model2.

From the above results, it was easy to see that, according to the BIC scores, the model with random effect in state only was the best. Therefore, we chose to use state as the only group variable in locations with random effect.

Selection Among Other Variables

Now, the base model only involves the location variable state with random effect.

First, from the **(mgstr EDA graph)**, we could see that there was a slightly negative relationship between the log(ppm) and the mgstr variable. Since the ppm was log-transformed, a slightly negative coefficient between log(ppm) and mgstr might be enlarged after exponentiation. Therefore, we considered the mgstr (dosage strength) as a quite important predictor and added it to the model to make comparisons based on the BIC score. By comparing with the base model, the model with the addition of mgstr as a fixed effect had much lower BIC score. Therefore, we decided to include mgstr in the model. Now, the model contained random effect of state and fixed effect of mgstr.

From the **EDA plots**, we also considered to add random slopes of mgstr by state to the model. However, by doing so, the BIC score of the model increased a lot compared to the model without random slope of mgstr by state. Therefore, mgstr was added to the model only as a fixed effect.

Second, from the **EDA Figure**, the log(ppm) seemed not to vary across bulk sizes, both in general and with-in sample states. Therefore, we considered bulk as an unimportant variable. However, adding bulk to the model slightly lower the BIC score. We preferred not to add random slopes of bulk by state to the model since within the sampled states, the difference of log(ppm) over bulk categories seemed not to change. Thus, we decided to add bulk to the model as fixed effect only, which now included random effect of state, and fixed effects of mgstr and bulk.

Third, we considered to add source and primary reason to the model. From the **EDA graphs**, within sampled states, the source factor seemed to have different effects on log(ppm) across different states. Therefore, we considered to add random slope of source by state and fixed effect source to the model, however, the BIC score of the model increased a lot after adding either fixed effect of source or both random effect and fixed effect of source to the model. Therefore, the source variable was excluded from the model. We also excluded the Primary Reason. From the **(Figure Primary)**, we could see that the response variable, log(ppm), did not change across different Primary Reason categories. And within the sampled states, the log(ppm) also did not change over different Primary Reason categories. Therefore, we would like to exclude Primary Reason from the final model. In addition, the model with primary reason as fixed effect or both random effect and fixed effect increased BIC values a lot, which confirmed our decision not to include it in the model.

Finally, based on the **EDA Figures**, it seemed that there was no obvious relationship between the log(ppm) and days ellapsed. However, we could observe slightly different slopes of days ellapsed within the sampled states. In the meantime, we thought as the local economy changed over time, the price of the morphine might also change over time as well, which also made us believe that the days ellapsed should be added to the model. And since each state had different economy development histories, the days ellapsed variable might have different relationships with log(ppm) in different states, which made us prefer to add random slope of days ellapsed by state to the model. Therefore, we added days ellapsed with random slope by state to the model along with its fixed effect. However, either adding fixed effect days ellapsed only or adding both random effect and fixed effect would increase the BIC score of the model. Therefore, we finally exclude this variable from the model.

Model Summary

From the above reasoning and testing results, our final model could be summarised as the following formula.

$$y_{ij} = \beta_{0,j}\ \text{State}_j + \beta_1\ \text{Bulk}_{i,j} + \beta_2\ \text{Dosage Strength}_{i,j} + \epsilon_{i,j}$$
$$\beta_{0,j} = \beta_0 + b_j;\ b_j \sim \text{Normal}(0, \tau^2)$$
$$\epsilon_{i.j} \sim \text{Normal}(0, \sigma^2)$$

In the above formula, b_j represents the random effects of `state`, and $\epsilon_{i,j}$ is the error term.