IDS702 Team Yellow Project II

Anna Dai, Athena Liu, Dauren Bizhanov, Himangshu Raj Bhantana, Moritz Wilksch

October 16, 2021

Summary

Introduction

Data

The data set used for the analysis is the subset with methadone as an active ingredient. It contains six variables: the outcome variable `ppm`, `state`, `USA_region`, `source`, `mgstr` and `bulk_purchase`. The data set contains missing values in three variables of interest including the outcome variable. All incomplete cases are removed. Moreover, the factor variable `source` has high cardinality with few cases in certain factor levels. It has been decided to group some levels to have clearer picture in exploratory data analysis. So, all internet based sources are grouped into single level named "Internet" and values such as "None", "N/A" are grouped into "No Input" category. The variable `mgstr` has six unique values and numeric data type. After initial inspection `mgstr` variable is transformed to a factor variable and 1mg, 2.5mg and 15mg cases are filtered as they have only one or two cases per value. The independent variable is highly skewed with a few outliers. After the check for typical methadone prices has been done, it has been decided to use the percentile method for ouliers removal with 95 percentile level as a cutoff.

!!! Not sure where to write about log scaling and normality assumption after outliers removal. !!!

During EDA it has been found out using box plots that price per milligram distributions are about the same for different source levels. But, there is a tendency for lower prices with higher dosage strength. Interestingly, different regions have different median prices per methadone milligram. The same situation across the states. Therefore, these variables may be potential candidate for a hierarchical model. Surprisingly, there is no much price difference by `bulk_purchase` variable.The only prominent interaction that has been observed is dosage streght by states. The trends are very different e.g. in Texas prices are getting lower with higher dosage strength and getting higher in Delaware.

- general description
  - e.g. free entry = dirty data?
- cleaning
  - e.g. cosolidating factors
  - removing unused levels (e.g. mgstr now only has 5, 10, 40)
- outlier removal (first 99, then 95)
  - alliviates need for log-scaling
- EDA
- Interactions

Model

- non-hierarchical linear model
- adding hierarchy

Conclusion