IDS702 Team Yellow Project II

Anna Dai (Presenter) Athena Liu (Checker) Dauren Bizhanov (Writer)
Himangshu Raj Bhantana (Coordinator) Moritz Wilksch (Programmer)

October 25, 2021

Summary

This study explores factors associated with the prices of methadone on the black market with a multilevel linear regression model. The response variable in our study is the price per milligram (`ppm`) variable. In combination, the linear regression models and the stepwise regression process suggested that `mgstr`, `source`, and `bulk_purchase` as predictor variables. `USA_regions` and `state` are suitable as hierarchical predictor variables. Our final model contains `mgstr`, `source`, and `bulk_purchase` as fixed effects, and `USA_regions` and `state` as random intercept variables. The fixed effects suggested that there is a tendency for lower prices with higher dosage strength as well as with bulk purchase. The hierarchical variables suggested that methadone prices could vary by region within the United States. For example, the price per milligram (`ppm`) in the South tends to be higher than in other regions. There are three significant states: California, Arizona, and Tennessee, in which California and Arizona seem to have the cheapest methadone prices, while Tennessee is paying significantly more compared to all the other states.

Introduction

Prescription opioid diversion and misuse are major public health issues. Street pricing reflects medication availability, demand, and potential abuse. However, such information can be challenging to obtain, and in an age of Internet-based social networks, crowdsourcing seems to be an effective solution. Nevertheless, for our study, we use data provided by StreetRx. StreetRx is a web-based citizen reporting tool that collects real-time street price data on diverted pharmaceutical medicines. Based on crowdsourcing ideas for public health surveillance, users can anonymously report drug prices they paid or heard were paid for diverted prescription drugs on the website. This study utilizes the product- and geographically-specific data on the drug Methadone from StreetRx. Methadone is an opioid class medication that helps to reduce drug withdrawal symptoms for other narcotic drugs. Unfortunately, methadone itself is also addictive. Knowing the price of methadone on the street would be crucial to preventing drug abuse, misuse, and the diversion of prescription drugs. Thus, this study focuses on exploring factors that influence the price of methadone per milligram. We aim to utilize a multilevel model to study characteristics associated with the price per mg of methadone, allowing for potential clustering by area and examining variability in pricing by region.

Data

The data set used for the analysis is the subset with methadone as an active ingredient. It contains 13 variables, out of which we are interested in six. Price per milligram (`ppm`) is our response variable, `source`, `mgstr`, `bulk_purchase` are our predictor variables, and `state` and `USA_region` are candidates for hierarchical levels. We removed the variable `form_temp` because Methadone is only available in pill form in the data set. In addition, the data set contains missing values in two variables of interest,`mgstr` and `ppm`, which is outcome variable. Therefore, the exploratory data analysis begins with data cleaning in eliminating missing data values from the data set.

The original factor variable `source` has high cardinality with few cases in certain factor levels. Therefore, we decided to group some levels to have a clearer picture in the exploratory data analysis. All internet-based sources, such as the different URLs, "Internet Pharmacy", and "Google" are grouped into a single level named "Internet", and values such as "None" and "N/A" are grouped into the "No Input" category. Moreover, all entries with missing `ppm` are removed, which also eliminated rows with missing `mgstr` values. The variable `mgstr` has six unique values and numeric data types. After an initial inspection, the `mgstr` variable is transformed to a factor variable, and 1mg, 2.5mg, and 15mg cases are filtered as they have only one or two entries per value, leaving only 5mg, 10mg, and 40mg. Usually Methadone pills come in 5mg or 10mg doses, while the 40mg pills are not FDA-approved and thus only appear on the black market.

Table 1: mgstr frequency

mgstr	1	2.5	5	10	15	40
count	2	1	781	3047	1	351

The distribution of the response variable `ppm` is highly right-skewed and contains quite a few outliers. The log transformation of `ppm` helps to reduce the skewness in the distribution. Thus, we would like to examine further the performance of log-transformed `ppm` in the modeling phase. Another potential factor that could influence the final model is the outliers within`ppm`. According to a research article by Surratt et al. (2013), the median price for methadone on the black market is $1.00 in 2013. Therefore, it is concerning that our data set contains several outliers that are up to 40 times greater than the median street price. Based on this data, we use the percentile method for outliers removal with a 95 percentile level (which works out to a $2 `ppm`) as a cutoff. Overall, 181 data points were removed from the data set, which corresponds to around 4% of the total data.

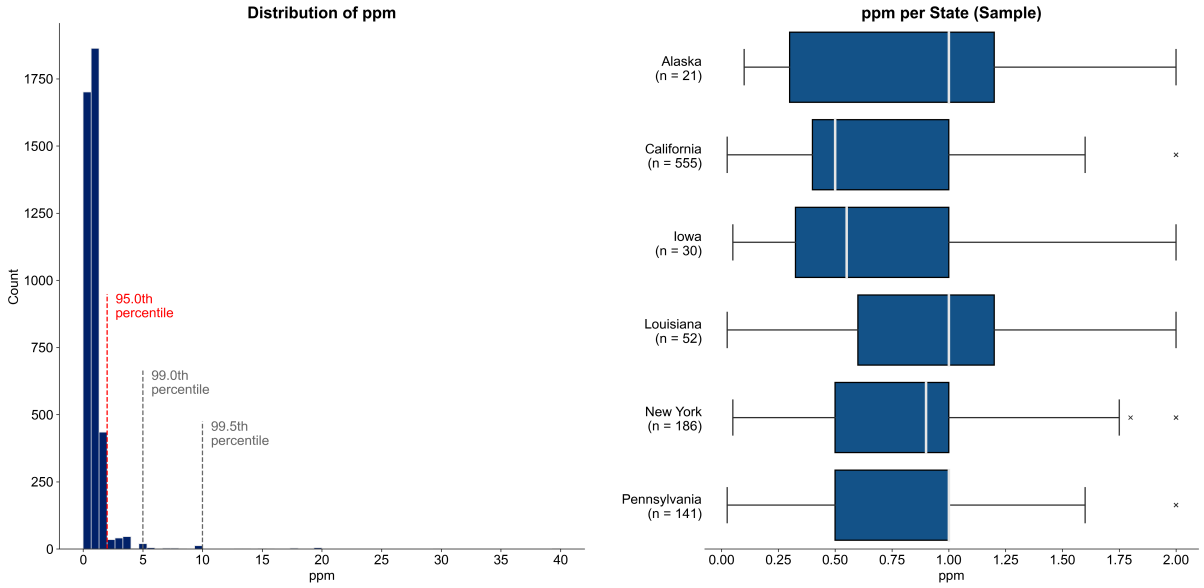

Figure 1: ppm and ppm by state distributions

During EDA, we found out (using box plots) that the 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 holds across the states which can be seen in the plot above which shows a subset of all states. Therefore, these variables may be potential candidates for a hierarchical level and as different states contain different number of observations, a hierarchical model may be better choice than multivariate linear regression with `state` and `USA_regions`

as categorical variables. Surprisingly, there is not much price difference in `ppm` by `bulk_purchase` variable.

As for the interaction observed between dosage strength and state, except for Texas and Delaware, the trend of `ppm` by dosage strength remains consistent across states. Similarly, the trend of `ppm` by `source` and `bulk_purchase` also do not appear to vary across states or regions. Therefore, for modeling, we decide to explore varying intercepts as opposed to random slopes.

Model

To fit the hierarchical linear regression model, we start modeling using a regular linear regression by defining a null model and full model to use in a stepwise regression process in order to build a parsimonious model. The null model contains only an intercept, whereas the full model contains all relevant non-location variables from the data set without interactions as predictor variables. For the stepwise process, we use the AIC as a decision criterion although using the BIC instead yields similar results.

Based on the original data (containing all outliers), the resulting model violated the normality assumption. To fix that we tried to log transform the `ppm` variable. Unfortunately, this does not help to meet normality assumption and makes residuals even less normal than before the transformation. Afterwards, we took the original `ppm` variable and removed outliers past the 95th percentile, which helps mitigate severe violation of the normality assumption.

Our stepwise model turn out to be the full model containing factorized `mgstr`, `source` and `bulk_purchase` as variables. All levels of `mgstr` are significant compared to their respective baselines: "5 mg" as well as "Bulk Purchase" significantly differs from "Not Bulk". Only the "No Input" source level is not significant compared to its baseline, "Heard it". Using this foundational model, we checked potential interactions including: `source:fac_mgstr`, `source:bulk`, and `fac_mgstr:bulk`. In order to do this, we employ the ANOVA F-test to test our original stepwise model against the stepwise model plus interactions separately and find that none of them significantly improve the model performance.

Given that the data set contains two naturally hierarchical variables `USA_region` and `state` and both of them are promising according to our EDA, we fit three random intercept hierarchical linear regression models including all the variables from the step model. We have the choice of using either `USA_region`, `state`, or both as hierarchical levels. The model using only `USA_region` has a noticeably higher AIC score, whereas `state`-only and the `state` + `USA_region` models have similar AIC scores (the model with both levels having the lowest AIC of 4756.9).

	AIC
region	4783.7
state	4760.2
state and region	4756.9

Random Intercept Hierarchical models AIC

To choose between these two, we use an ANOVA test. Its result suggests that the model using both `state` and `USA_region` is significantly better than the model that only models random intercepts by `state`. Note that the absolute AIC values between tables differ slightly due to ML/REML refitting.

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
state_only	9.00	4717.62	4774.26	-2349.81	4699.62			
state_and_region	10.00	4715.35	4778.29	-2347.68	4695.35	4.27	1	0.0389

F-Test for one random intercept vs two random intercepts models

Following the results of the ANOVA test, our final model contains `mgstr`, `source`, `bulk_purchase` and two hierarchical variables `state`, `USA_region`.

$$\text{ppm}_i = (\beta_0 + \gamma_{0j} + \psi_{0k}) + \sum_{m=2}^{3} \beta_{1m} \cdot \mathbb{I}[\text{mgstr}_{ijk} = m] + \beta_2 \cdot \text{bulk\_purchase}_{ijk} + \sum_{m=2}^{4} \beta_{3m} \cdot \mathbb{I}[\text{source}_{ijk} = m] + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2), i = 1, ..., n; j = 1, ..., J; k = 1, ..., K$$
$$\gamma_{0j} \sim \mathcal{N}(0, \tau_0^2), i = 1, ..., n; j = 1, ..., J$$
$$\psi_{0k} \sim \mathcal{N}(0, \varsigma_0^2), i = 1, ..., n; k = 1, ..., K$$

As mentioned before, our final model satisfies all linear regression assumptions. However, we acknowledge the fitted versus residuals plot has weird artifacts. This is most likely due to the fact that the data set contains only categorical variables.

Table 2: Fixed effects of the hierarchical linear regression model

	Estimate	Std. Error	t value	p value	Lower Bound	Upper Bound
(Intercept)	1.1427	0.0330	34.6471	0.0000	1.0789	1.2131
fac_mgstr10	-0.2927	0.0182	-16.0619	0.0000	-0.3283	-0.2569
fac_mgstr40	-0.4623	0.0291	-15.8731	0.0000	-0.5191	-0.4049
bulk_purchaseBulk	-0.0771	0.0172	-4.4743	0.0000	-0.1110	-0.0435
sourceInternet	-0.1190	0.0335	-3.5527	0.0004	-0.1843	-0.0531
sourceNo Input	-0.0349	0.0190	-1.8392	0.0660	-0.0721	0.0022
sourcePersonal	-0.0579	0.0187	-3.0919	0.0020	-0.0947	-0.0213

All the fixed effects in the final model are significant. The only exception is the "No Input" category of the `source` variable in comparison to its baseline "Heard it". As all fixed effects coefficients are negative, we can conclude that the highest price is predicted when all fixed effects are at their respective baseline levels, which represents a 5mg dose not purchased in bulk and the source of which has been reported as "Heard It".

As the final model contains many variables, it is easier to understand the model with a prediction plot which incorporates all the variables except `bulk_purchase` which will only move the graph along the y-axis. It is evident that `ppm` is lower for higher dosages as well as Internet-based sources.

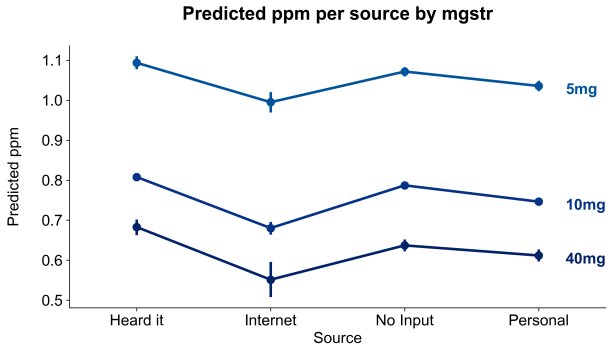

Figure 2: Prediction plot

Looking at the summary table below, the random effects' standard deviation for `state` variable is 0.057 and `USA_regions` is 0.0456 which are 10.6% and 8.5% of the whole variance, respectively. This indicates while these variables do explain some variance, they leave a lot of unexplained variance.

Table 3: Variance of the random effects

Groups	Name	Variance	Std.Dev.
state	(Intercept)	0.0033	0.0570
USA_region	(Intercept)	0.0021	0.0456
Residual		0.1878	0.4334

Examining the plot of random intercepts by region, we have only one region with a random intercept that is significantly different from the grand mean. The price per milligram in the South tends to be higher than in other regions. Similarly, there are three states with significantly different random intercepts: California, Arizona and Tennessee. Specifically, `ppm` in California and Arizona tend to be lower than average, while prices in Tennessee are higher than average.

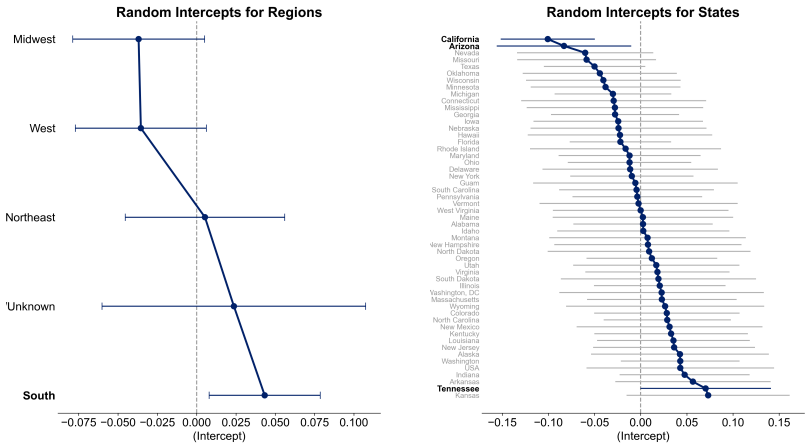

Figure 3: Random intercept by state and region

Conclusion

In this study, we studied factors associated with the price of methadone in the United States black market. We used a multilevel hierarchical model with random intercepts to analyze how product characteristics and geographical data could affect the price per milligram of methadone. The model used both `USA_regions` and `state` as hierarchical levels, and `mgstr`, `source`, and `bulk_purchase` as fixed effects. Geographically, the model suggests that methadone is the most expensive in the southern regions of the United States. Similar to this finding, the other hierarchical variable states confirmed this trend because the southern state Tennessee has the significantly most expensive methadone compared to the other states. On the other hand, California and Arizona have the least costly methadone nationwide. Moreover, we found the price of methadone varies with different dose strength, obtained source, and bulk purchase. The price of methadone tends to be more expensive when dose strength is weak and less expensive when dose strength is strong. Also, methadone tends to be less costly when purchased in bulk. Lastly, Internet and personal purchases are associated with cheaper methadone than the prices people heard of and reported.

There are multiple limitations of our analysis. First, StreetRx collected the data set through crowdsourcing, where users can enter methadone prices and related data. Thus, users could make mistakes while entering the values or intentionally misreport prices. For example, several of the entries for methadone prices are 40

times higher than the median price of methadone. We tried to remedy this issue by employing a common outlier removal technique, but have to acknowledge that all reported prices are subject to human mistakes and biases. Another limitation of the data set is the imbalanced distribution of cases between the categories. For example, some categories in sources and states contain only one or two data points. Future research could aim to collect a more comprehensive data set with a sufficient number of samples for all subgroups. Moreover, in the future, the relationship between methadone prices and other opioid drug prices could be studied. While methadone is an opioid used to relieve drug withdrawal symptoms, it is also addictive. Therefore, it will be crucial to understand if using other types of opioids influences the pricing and, therefore, the abuse and misuse of methadone. This could help to prevent abuse and misuse of multiple drugs in a region.

Citations

1. H.Surratt et al. 2013. Street prices of prescription opioids diverted to the illicit market: data from a national surveillance program. https://doi.org/10.1016/j.jpain.2013.01.455