IDS 702 Group Project 2: StreetRx

Peining Yang (Checker), Sarwari Das (Coordinator), Michelle Van (Presenter),

John Owusu Duah (Programmer), Satvik Kishore (Writer)

Summary

In this report, we analyzed data regarding street prices of Tramadol, a pain relief medication, collected through a crowd-sourcing website StreetRx. We investigated the factors that are associated with the price of Tramadol, as well as whether the difference in location contributed to the variation in price. Results showed that the drug's dosage is statistically significant in association with the price, with random variations based on the state and US region where the information is reported.

Introduction

StreetRx (streetrx.com) is a website where users can self report information regarding street prices of various pharmaceutical substances. It allows users to anonymously input prices they paid or heard about for prescription drugs, which provides insights about the black market and enhances efforts of public health surveillance. In this assignment, we will analyze data collected by the *StreetRx* website on Tramadol, a narcotic used to treat mild to severe pain. We will fit a hierarchical model in order to investigate the factors influencing the price/mg of Tramadol, accounting for the potential clustering by state and US regions.

Data

The dataset we used contains 5 response variables, which are *state* and *USA_region*: factor variables of the state and US region in which the information is reported from, *source*: which is a factor variable of the source of information, *mgstr*: which is a factor variable of the dosage strength in mg of the units purchased and *bulk_purchase*: which is a binary variable indicating whether the drug was purchased in bulk or not. The response variable of our analysis is *ppm*, which is a numeric variable for price per mg.

Originally, the *source* variable contained multiple websites that user entered. For ease of analysis, we grouped all sources which are websites into the level "Internet". In addition, we also had a *form_temp* variable, which is the formulation of the drug of either pill/table or patch. However, there were only one observation that had the form of patch, which is insufficient for our analysis. Therefore, we decided to exclude the *form_temp* variable. Additionaly, we had 2 variables for which dosage was 1, which is unlikely for a drug like Tramadol; we delete these rows as well.

Exploratory Data Analysis

We first investigated our outcome of interest, which is the *ppm* variable. After plotting a histogram, we observed that the distribution of *ppm* is extremely right skewed. This prompted us to perform a log transformation, and the results showed a much more normal distribution. We will proceed with *log(ppm)* as our response variable.

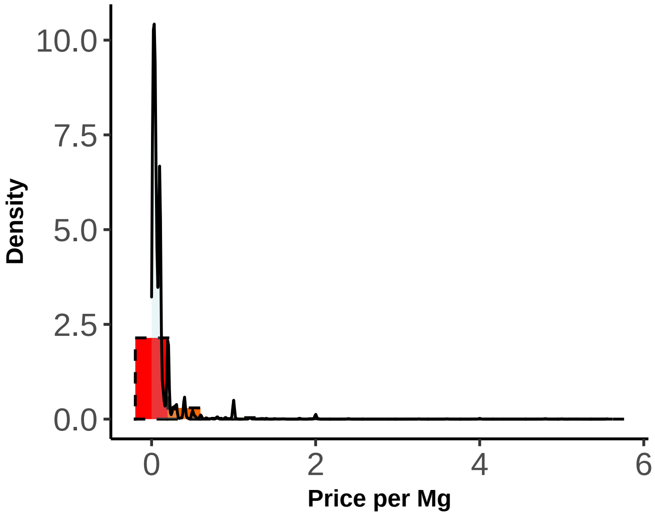


Figure 1: Distribution of Price per mg

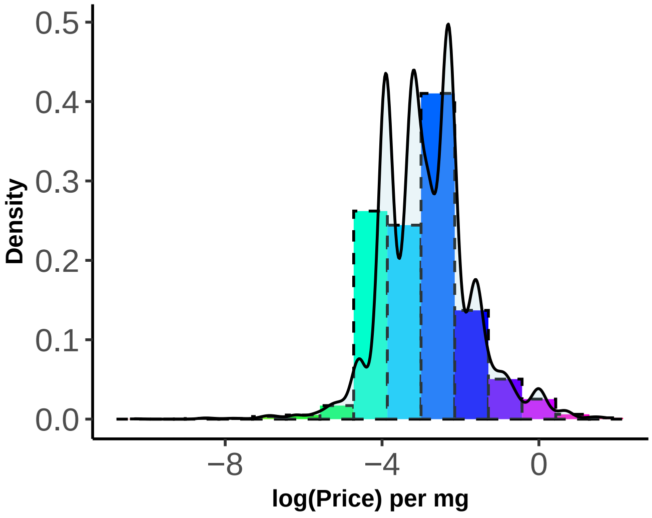


Figure 2: Distribution of Log(Price) per mg

Since we are interested in exploring the heterogeneity of the price of Tramadol by location, the boxplot below shows the log(price) of Tramadol by a subset of states. Results show that there is indeed variation between locations. In addition, since states are nested within US regions, a boxplot of Tramadol prices by region did also show variation. Therefore, we will include both states and regions as random effects of our model.

From examining the price variable, we discovered that the minimum value is 0.000033 while the maximum value is 5.56. This means that there are possible outliers in the model, which is also depicted in the boxplot below. However, due to the nature of the data set, there is not a suitable way to verify whether an observation is an outlier or not. Therefore, we will keep this in mind during our analysis and have decided not to remove the outliers.

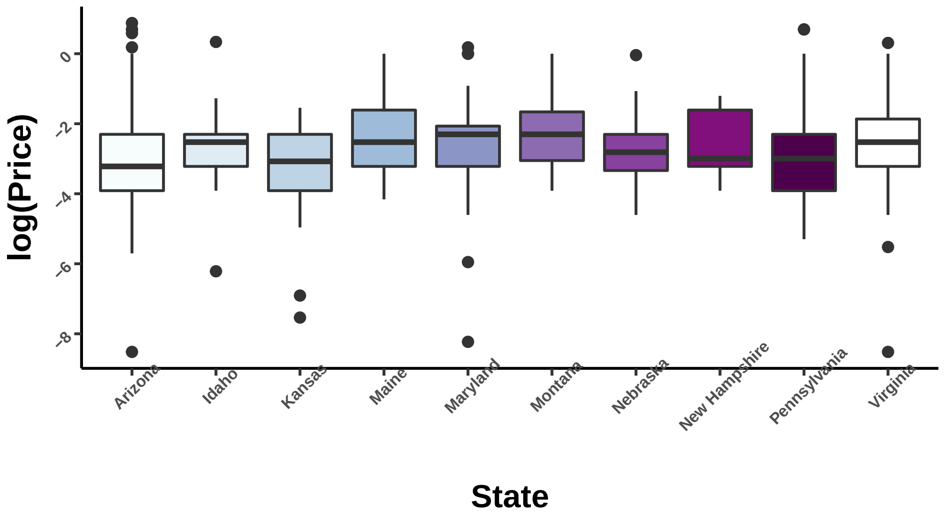
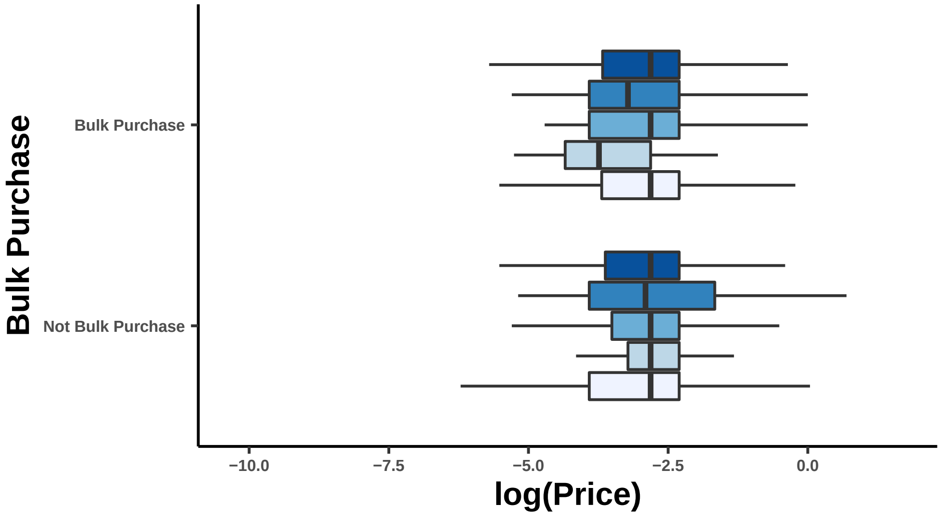


Figure 3: log(Price) of Tramadol by State

Lastly, we explored any potential interaction effects that should be included in our model. As seen from the figure below, we see that the price of Tramadol for those that were bulk purchases are influenced by the source of the information. No other interaction effects that we experimented with were deemed significant. Therefore, we will include this interaction effect in our model.

Figure 4: Interaction between Bulk Purchase and Source



Hierarchical Linear Regression Model

Our initial model included fixed effects of *mgstr*, *bulk_purchase* and *source*. As mentioned in our exploratory data analysis, we expect there to be an interaction effect between *bulk_purchase* and *source*. After fitting a simple linear regression model with and without the interaction, an analysis of variance (ANOVA) test showed a p-value of 0.027, which is below the 0.05 threshold. Therefore, we will include this term in our final model. We also conducted a backward AIC with the full model being a linear model with all possible main effects and interacts. AIC confirmed the results of our EDA, and the final model only had the main effects and the one interaction in it. Next, we wanted to determine whether varying slopes between *bulk_purchase* and *state* and *source* and *state* are statistically significant. After fitting a hierarchical linear regression model for both and performing an ANOVA test with the simple linear regression model mentioned before, results showed p-values of 0.228 and 0.948, which are well above the threshold. We will proceed without any varying slopes in our model.

As seen from the figures in our EDA, we expect the price of Tramadol to varying by location. Since geographically, states are nested within US regions, we then determined whether we should include only *state* as the random effect or include both *state* and *USA_region*. After fitting two hierarchical linear regression models with and without *USA_region* as the varying intercept in addition to state and the fixed effects mentioned above, an ANOVA test produced a p-value of 0.01. In addition, the model with two varying intercepts has an AIC value of 14934.55 while the model with only one varying intercept has an AIC value of 14941.98. Therefore, our final model will include both state and US regions as the random effects. The mathematical representation and model output of our final model is shown below.
$log(Price_{ijk}) = (\beta_0 + \gamma_{0k} + \gamma_{0jk}) + \beta_1 bulk\_purchase_i + \sum_{a=1}^{7} \beta_{2a}[mgstr_i = a] + \sum_{b=1}^{5} \beta_{3b}[source_i = b] + \sum_{b=1}^{5} \beta_{4b} bulk\_purchase_i : [source_i = b] + \epsilon_{ijk}; i = 1, ..., n_j; j = 1, ..., J$
where a takes on different levels of the dosage variable and b takes on different levels of the source variable.

$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$
$(\gamma_{0k}, \gamma_{0jk}) \sim \mathcal{N}_{\in}(\mathbf{0}, \Sigma)$

Table 1: Hierarchical Linear Regression Model Output

Fixed Effects					
	Estimate	Std. Error	t value	2.5 %	97.5 %
(Intercept)	0.660	0.765	0.862	-0.802	2.186
mgstr_f37.5	-3.047	0.761	-4.004	-4.537	-1.556
mgstr_f50	-3.551	0.759	-4.681	-5.034	-2.064
mgstr_f100	-4.172	0.761	-5.482	-5.664	-2.684
mgstr_f150	-4.517	0.772	-5.850	-6.037	-3.013
mgstr_f200	-4.434	0.771	-5.755	-5.943	-2.925
mgstr_f300	-5.030	0.773	-6.508	-6.539	-3.513
bulk_purchaseBulk Purchase	0.018	0.071	0.248	-0.122	0.156
source_newInternet	0.021	0.100	0.205	-0.192	0.208
source_newHeard it	0.000	0.054	-0.002	-0.106	0.106
source_newInternet Pharmacy	0.288	0.123	2.334	0.047	0.531
source_newPersonal	-0.023	0.040	-0.579	-0.101	0.054
bulk_purchaseBulk Purchase:source_newInternet	-0.221	0.191	-1.155	-0.680	0.133
bulk_purchaseBulk Purchase:source_newHeard it	0.008	0.122	0.065	-0.231	0.248
bulk_purchaseBulk Purchase:source_newInternet Pharmacy	-0.364	0.273	-1.332	-0.898	0.172
bulk_purchaseBulk Purchase:source_newPersonal	0.078	0.095	0.820	-0.107	0.265

Random Effects		
Groups	Variance	Std.Dev
state	0.0002	0.013
USA_region	0.0457	0.214
Residual	1.1500	1.077

Model Assumptions

With our final model, we performed various diagnostics to assess model assumptions. Since, most variables in our dataset are discrete, we are concerned mostly about the independence and normality of residuals assumption. On plotting residual vs fitted value plots (below), we see that the independence assumption can be improved; while there is randomness in the Y axis, points are clustered on the X axis, probably due to missing variables. Further variance is not constant. The plot of residuals of the final model also showed that many points deviated from the 45 degree line. The QQ plot of the model is shown below. This indicates that there is a violation of the normality assumption with our model. We have already performed a log transformation on the response variable. In addition, we also try to remove the aforementioned outliers in hopes to fix this issue yet it showed minimal improvement. The removal of the outliers only marginally changed the model, indicating that these observations are not high influential points. Given the nature of the data, we decided that it was best to proceed. The Variance Inflation Factors (VIF) of the final model are all below 10, implying that there are no issues of multicollinearity.

Figure 5: Residuals vs Fitted Values

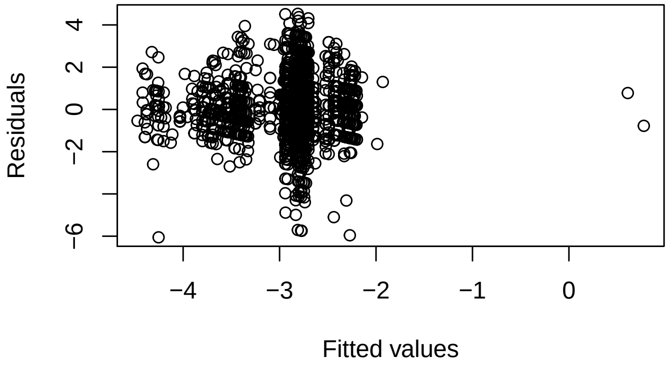
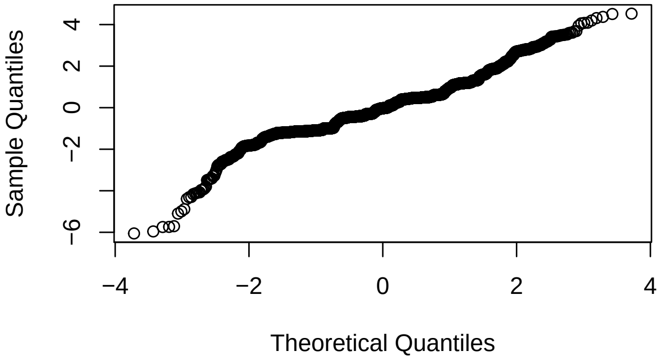


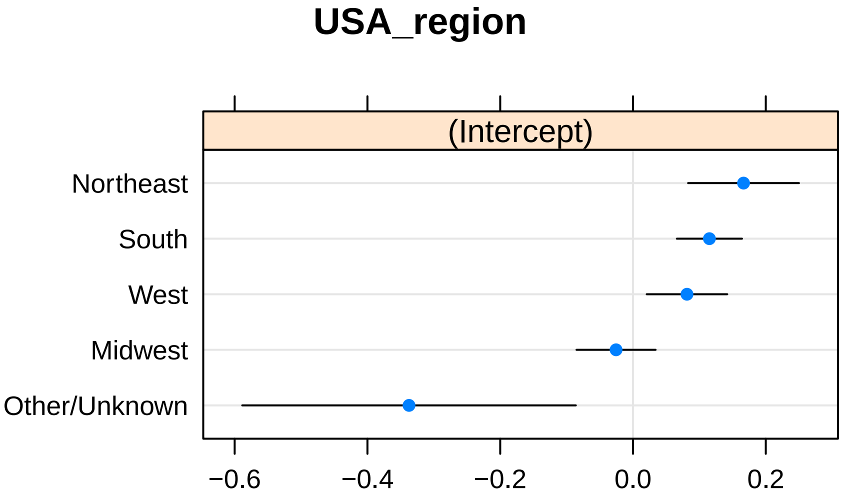
Figure 6: Normal Q−Q plot



Model Interpretation

The baseline of our final model is Tramadol of 1mg dosage, non-bulk purchase where the source of the information comes from Others and we expect the price per mg to be $e^{0.66} = \$1.93$. When the purchase is in bulk, keeping all else constant, the price per mg of Tramadol is expected to increase by a multiplicative effect of $e^{0.018} = 1.02$, which is about a 2% increase. All dosage indicators are significant and higher dosages are associated with lower prices. For example, compared to the baseline of 37.5mg, a dosage of 300mg (keeping all else constant) leads to a decrease in price by 1- $e^{-1.98} = 0.13$, which is about 87%.

The estimated standard error for state is 0.013, which describes the across state variation attributed to the random intercept. For regions, the standard deviation is 0.21. This implies that ppm of Tramadol is varying more by region, than it is by state. Further since state and region are nested in nature, we expect the overall geographical variation in prices to get distributed between the two. The estimated standard error of the residual of the model is 1.14, which describes the within-state/region or the remaining unexplained variation. As seen from the figure below, our model is not statistically significant for th Midwest region but it is for the remaining regions.



Limitations

In the process of our analysis, we noted several limitations that could be improved upon with future work. First, StreetRx is a crowd sourcing website where users can input any information they want. This means that there are no methods to check for the accuracy of our data. For example, the dataset contained 7 observations where the prices of Tramadol were missing yet other information such as state and USA_region had an input. Second, we also removed the variable indicating the form of the drug as there were only 1 observation for “patch” while the remainder were all “pill/tablet”. Intuitively, this should be an important variable in predicting the price of Tramadol yet was excluded from our analysis due to the lack of data. We also removed the date variable indicating the time when the drug was purchased. This could also be a significant variable as the price of Tramadol could be changing over time. Lastly, we noted that several model assumptions of our final model were not met. As mentioned before, this is likely due to the fact that we dropped several potentially significant variables and were only working with a subset of the data.

Conclusion

The opioid crisis in the United States is an urgent issue that desperately needs a solution. Our analysis showed that there is a clustering effect of price of Tramadol that varies across US regions. In addition, we also identified the dosage of Tramadol and when the source of information is internet pharmacy to be influential factors for the price per mg of Tramadol. This analysis is crucial in gaining insights regarding the opaque black market of illegal prescription substances, which would hopefully contribute to the continued effort of resolving the opioid epidemic. The clustering of price by region would aid in creating policies that are more suitable for local communities. To build upon our current analysis, we could potentially investigate whether a state’s legalization status of marijuana contributes to the street price of other narcotics. Marijuana, which currently is often used medically to treat chronic pains, could be a better alternative to addictive pain-relief substances.