Street Price of Lorazepam

Team: Erika Fox (Programmer), Minjung Lee (Presenter), Preet Khowaja (Checker), Raza Lamb (Coordinator), Robert Wan (Writer)

10/24/2021

Summary

Using a hierarchical linear regression model and data reported by citizens anonymously, we investigated the factors that impact the per mg street price of Lorazepam in the United States. We found that the street price is most impacted by the dosage strength per unit, with stronger pills generally costing less per mg. The street prices in most states are similar, although 6 states have street prices that are statistically different from the national average.

Introduction

Some studies place the size of the street market of diverted pharmaceutical substances at around $400 billion per year. However, it is an opaque market where player activities are actively concealed and pricing dynamics are often not well understood. A good understanding of the prices of drugs on the street can be helpful for public health officials as they seek to combat substance abuse issues.

In this study, we investigated on the street price of Lorazepam. Lorazepam is a benzodiazepine that is used to treat anxiety disorders. Abuse of Lorazepam has been associated with a number of health issues, including impaired muscular coordination and memory loss. In the United States, Lorazepam is a Schedule IV controlled substance that can be legally acquired only through prescriptions, which creates a demand on the street for the drug. Specifically, we sought to understand what factors are associated with the street price of Lorazepam in the United States, and how street price varies by state.

Data

Data Source

For this analysis, we got our data from StreetRx (streetrx.com). StreetRx is a web-based citizen reporting tool that allows people to anonymously report the street price of diverted pharmaceutical substances they are aware of or have purchased. Using user-generated data allows us to get insights about an otherwise opaque market. However, we have to be weary about potentially incorrect values in the data set because the data points have not been double-checked.

Code Book

- ppm: price per mg or Lorazepam, in USD
- $state$: the US state that the user entered
- USA_region: the region of the US based on the state
- $source$: the source that the user got the price
- $form$: the formulation of the drug
- $mgstr$: dosage strength of the drug purchased, in mg / unit
- $bulk\_purchase$: 1 if the purchase was of 10+ units, 0 otherwise

Data Cleaning

Since the data was reported by individual users instead of collected by professionals, some data points might be incorrect. We found a few problems with the data set and took measures to address them, which are listed below.

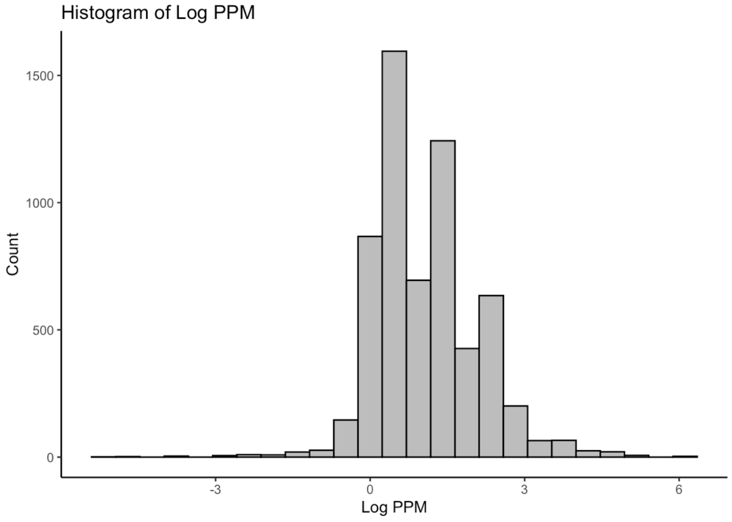
1. **Dropped observations with ppm above $100/mg**: around 20 observations have prices of over $100/mg,

with a few as high as $400/mg. The average price of Lorazepam is around $1/mg in pharmacies, so prices in the hundreds are likely due to mistakes in reporting. However, since there is not an authoritative source of the real price of Lorazepam, we wanted to be conservative in deciding which prices were errorneous. By dropping observations with ppm above $100/mg, we dropped 0.3% of observations, which we believed strike a good balance between keeping the data clean and retaining enough information.

2. **Dropped observations with strength above 2.5mg/unit**: two observations had strength (dosage strength per unit) above 2.5mg/unit. Lorazepam is offered in strengths from 0.5mg/unit to 2.5mg/unit. Since only two observations had strength above 2.5mg/unit, we were comfortable with removing both observations.
3. **Dropped observations where the $source$ have fewer than 5 reported prices of Lorazepam**: because users were allowed to enter their source in free text form, some sources were web urls or had unique names. This resulted in many distinct values of the $source$ variables with five or fewer observations. A total of 28 observations were like this. These observations would not be useful for building the model, and we dropped them because they made up less than 0.5% of the data set.
4. **Transformed $mgstr$ into a factor variable**: because there are only 4 distinct values of $mgstr$, we transformed it into a factor variable, named $mgstr_factor$.
5. **Combined states / territories with 10 or fewer observations into one group**: American Samoa, Washington DC, Virgin Islands, and Wyoming had 10 or fewer observations. We combined these four states / territories into one group to avoid problems caused by small sample sizes.

Exploratory Data Analysis

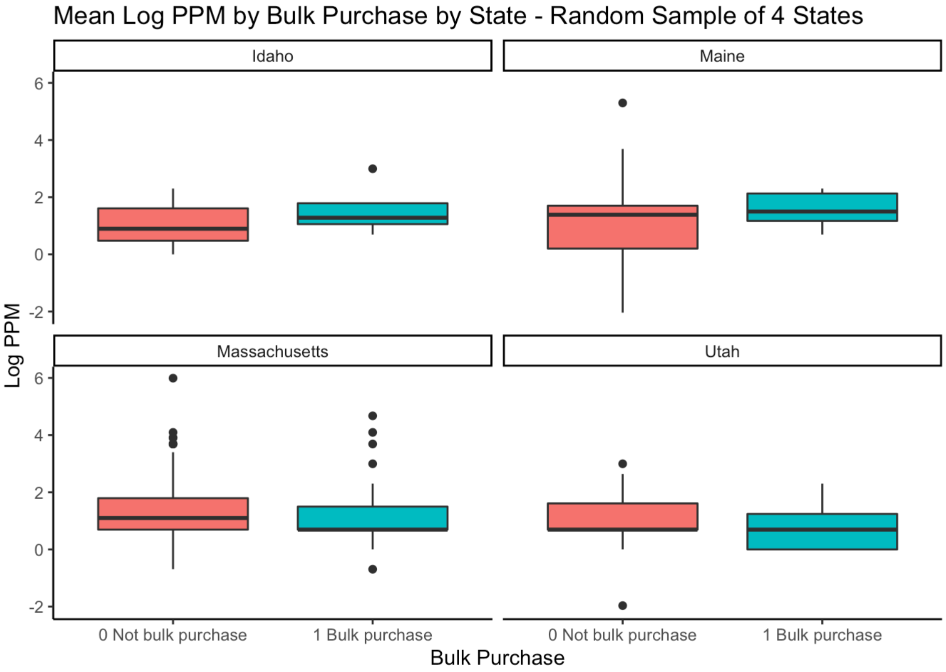
The variable we would like to use as the response variable of the model, ppm, is a continuous variable, so it would make most sense to use a linear regression. As such, we first checked if ppm follows a normal distribution. It turns out that having removed observations with ppm > $100, the distribution of ppm was still not normal. On the other hand, the natural log of ppm, named ppm_log, followed a normal distribution, as shown in Figure 1. We decided to use it as the response variable instead.



Having finalized the response variable to use, we moved on to investigate potential predictor variables for the model. Our data set includes two location variables, $state$ and USA_region. This prompted us to consider creating a hierarchical model group by geography. We explored weather the average ppm_log differed by those two variables, We found that ppm_log differed substantially by $state$, but there were no discernible differences between each USA_region. As such, we believed it would be helpful to group by $state$ in the model but not as helpful to group by USA_region. We would take another look at the hierarchies when we construct the model.

We then looked at other variables in the data set, namely $source$, $form$, $mgstr\_factor$, and $bulk\_purchase$. All Lorazepams in the data set come in the form of pill/tablet, so we discarded this variable for modeling. The remaining three variables all seemed be correlated with ppm_log to a degree, so we decided to include all of them in the null model we would build.

In terms of interaction variables, we did not find many interactions to be interesting - for most interactions, the pattern of average ppm_log did not change between the groups created by the interactions. However, we found one interaction that could be worth to further investigate. As shown in Figure 2, the effect of $bulk\_purchcase$ appeared to differ by state. Lorazepam bought through bulk purchases had high prices in states like Idaho, whereas it cost less in states such as Maryland. This suggests that we could explore whether adding a varying slope by $bulk\_purchase$ would improve the model.



Model

The final model we selected was a hierarchical linear regression model with varying intercepts for each state. The fixed effects predictors were source, bulk purchase, and dosage strength as a factor variable, and there was no interaction variable. In this model, each state has its own intercept, which is an effect in addition to the common intercept. The formal representation of the model is:

$$y_{ij} = (\beta_0 + \gamma_{0j}) + \beta_1 source_{ij} + \beta_2 bulk\_purchase_{ij} + \beta_3 mgstr\_factor_{ij} + \epsilon_{ij}; \quad i = 1, \ldots, n_j; \quad j = 1, \ldots, 51$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$
$$\gamma_{0j} \sim N(0, \tau^2_{state})$$

Variable Selection

When we started model selection, we used the above final model as our base model. This model made a good base since it has no interactions and one varying intercept by the state variable, and we were sure that we would like to build a hierarchical model grouped at the state level.

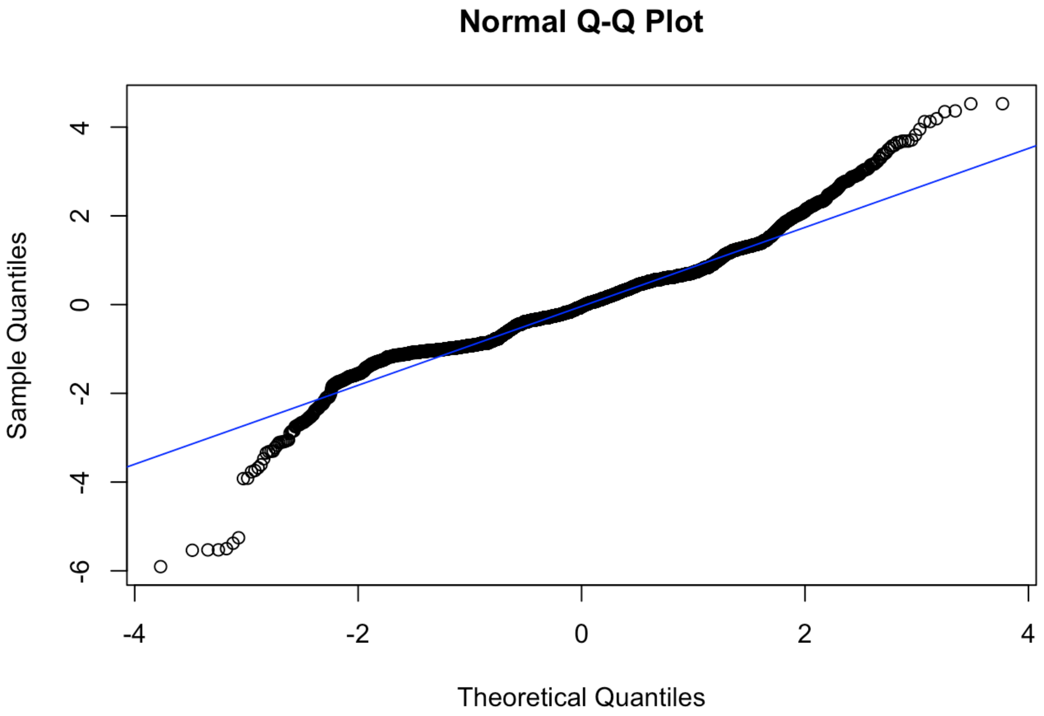
Even though we did not find any interesting interactions for fixed effects in the EDA, we added interactions terms to the model in order to ensure that we did not miss important predictors. We performed ANOVA tests to compare the base model and models with interaction terms. It turned out that no interaction term we attempted to add improved the model fit in a statistically significant way. As such, we decided to not include any interaction terms for fixed effected in the final model.

In the EDA, we found that the interaction between $bulk\_purchase$ and $state$ could be worth further investigations. As such, we added a varying slope for $bulk\_purchase$ grouped by $state$. However, we got singular fits when we tried to train the model regardless of the optimizer we used in the lmer() function. This indicated that the model likely suffered from overfit issues when we added the varying slope. The structure of the random effects we tried to model was too complex to be supported by the amount of data we had. That was entirely reasonable since we had a little over 6,000 observations in the data set and $state$ already split the data into 51 groups. Therefore, we resorted to using only a varying intercept for $state$ and no varying slope.

Model Asesssment

To assess the four assumptions of linear regression, we first looked at a scatterplot of residual vs fitted values. Residuals were evenly scattered around 0 and did not have any pattern with respect to fitted values. As such, the model passed the independence assumption and the equal variance assumption. We also did not notice any influential outliers from the plot. The linearity assumption does not apply to our model because all predictors are factor variables.

Things become trickier regarding the normality assumption. The x and y axes of the QQ plot below represent the theoretical and actual z-scores of each residual. Residuals with theoretical z-scores between -1.5 and 1.5 lie on the QQ line, which translates to roughly 88% of residuals following a normal distribution. Residuals with more extreme values tend to spread out more than what we would expect from a normal distribution. Since the data set still includes some extreme values for ppm, this is not unexpected - however, this would be worth diving deeper in future research and could possibly be resolved as we gather more data and become more confident on a reasonable range for the street price for Lorazepam.



Results and Interpretation

A table of model coefficients and a plot showing the random intercept of each state are included in the appendix. For fixed effects, the factor variable for dosage strength, $mgstr\_factor$, is the most influential predictor. Pills with higher dosages generally have a lower price per milligram than pills with lower dosages. Holding all else equal, compared to pills with a dosage strength of 0.5mg / unit, 1mg / unit pills cost 45.05% less per mg, and 2mg / unit pills cost 64.56% less per mg The highest dosage pills, with a strength of 2.5mg / unit, constitute only 1.7% of all observations, which suggests that there might be less supply of pills with such a high dosage. It also means that we would not be able to infer the ppm of these pills as accurately. As such, the ppm of 2.5mg / unit pills is slightly higher than the ppm of 2mg / unit pills - compared to 0.5mg / unit pills, 2.5mg / unit pills cost only 60.36% less per mg. At the 0.05 level, other fixed effects are not statistically significant. However, the coefficient of $bulk\_purchase$ is significant at the 0.1 level, and drugs purchased in bulk are 5.78% cheaper per mg, holding all else equal.

For random effects, the standard deviation of the varying intercept is 0.1159 while the standard deviation of the residuals is 0.9264. The varying intercept by state explains 12.51% of the variation of ppm, while the fixed effects explain the remaining variations in ppm within each state. The confidence intervals of the intercept for most states include 0, so the baseline price in most states are not significantly different from the average baseline price of the country. However, a few states have prices that are statistically different from the average price of the country - all else held equal, baseline prices in NJ, FL, and CA are much higher than the national average and baseline prices in MI, PA, and IA are much lower.

Conclusion

In this study, we found that the most influential factor of the street per mg price of Lorazepam is the dosage strength. For pills with strengths between 0.5mg / unit and 2mg / unit, higher strength translates to lower per mg prices. However, pills with the highest dosage strength, 2.5mg / unit, cost slightly more than pills with 2mg / unit, which is likely due to limited supply on the street. Purchasing Lorazepam in bulk does tend to lower the per mg price, yet the effect is not significant at the 0.05 level, and the magnitude of the effect does not appear to be scientifically meaningful. The location of the transaction does not play a meaningful role in determining the street price of Lorazepam. While a few states have statistically different baseline per mg prices than the national average baseline, most states have similar prices.

Limitations

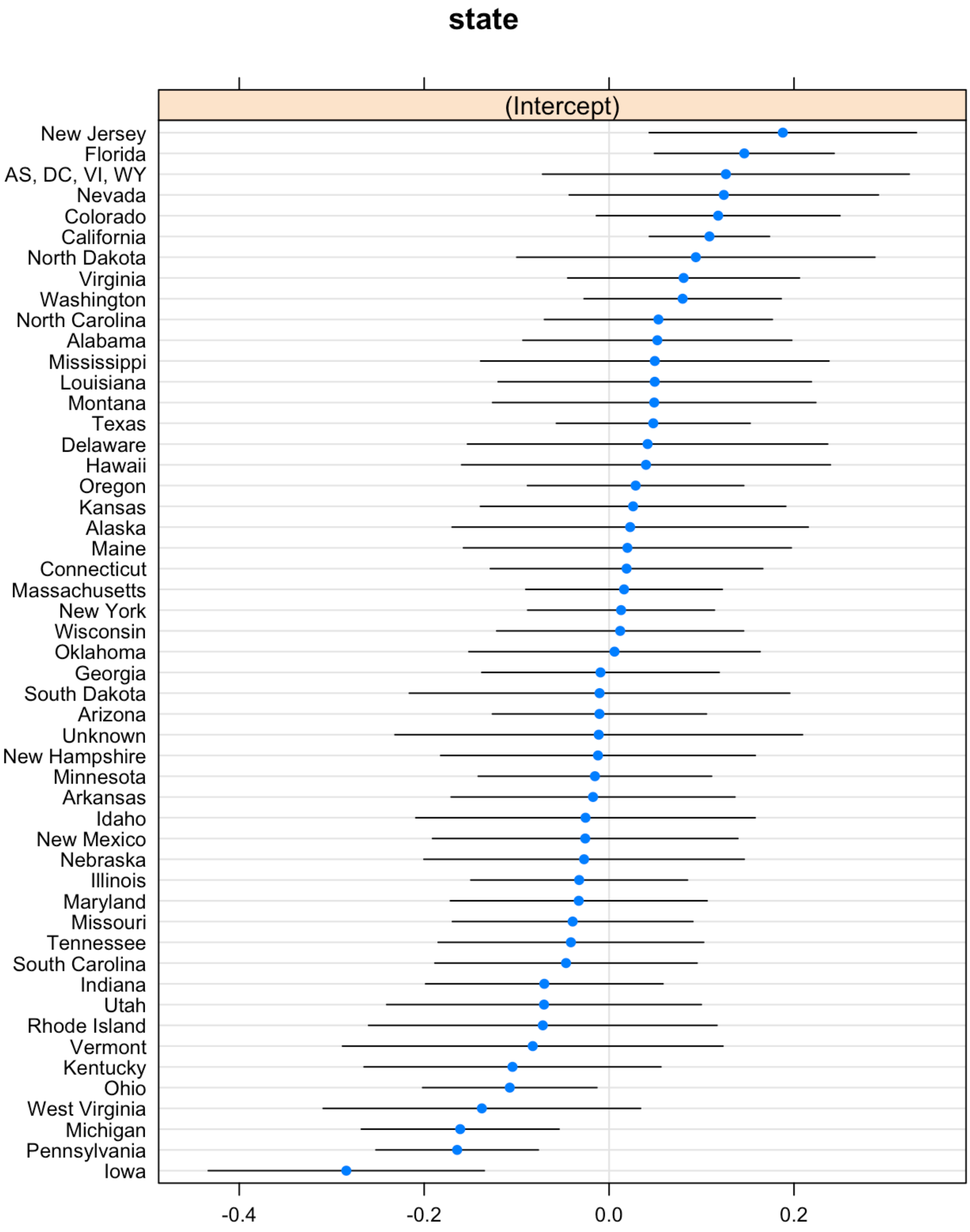
As previously mentioned, one of the biggest limitations of this analysis is the source of our data, in that all the information we have is user-generated and therefore unreliable. User-generated data tends to have many quality issues. Although we have dropped observations that we believed to be wrong, we could never be absolutely sure that the data is free of errors. This problem is exacerbated because we do not know the true range of the street price of Lorazepam. On one hand, we might have not dropped all observations with wrong prices. On the other hand, we might have dropped observations that actually had correct values, but they were too extreme that we thought they were incorrect.

Our study could also benefit from having more data. Some states have number of observations in the high tens. As mentioned in the modeling section, the model suffered from overfit issues when we tried to add varying slopes. If we had more data, we might be able to fit more complex models so that we can infer price more accurately. In addition, for this analysis, we assumed that the log of ppm has a normal distribution. However, the QQ plot showed that residuals on both ends deviated from a normal distribution. More data might better tell us what the real distribution of street price is so that we can fit an appropriate regression model.

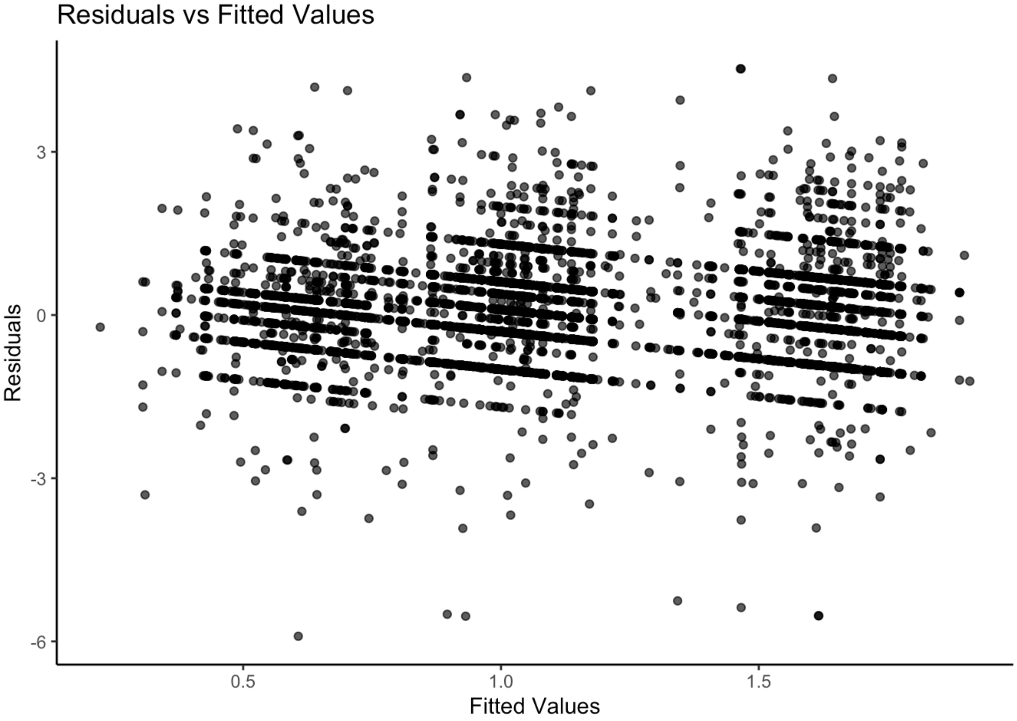
Appendix 1: Model Coefficients

	ppm log		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.63	1.57 – 1.69	**<0.001**
source [Heard it]	-0.06	-0.13 – 0.01	0.122
source [Internet]	-0.08	-0.22 – 0.05	0.232
source [Internet Pharmacy]	0.15	-0.02 – 0.33	0.083
source [Personal]	0.00	-0.05 – 0.06	0.879
bulk_purchase [1 Bulk purchase]	-0.06	-0.12 – 0.00	0.066
mgstr_factor [1]	-0.60	-0.65 – -0.55	**<0.001**
mgstr_factor [2]	-1.04	-1.11 – -0.97	**<0.001**
mgstr_factor [2.5]	-0.93	-1.11 – -0.74	**<0.001**
Random Effects			
σ^2	0.86		
$\tau_{00\ \mathrm{state}}$	0.01		
ICC	0.02		
N_{state}	51		
Observations	6075		
Marginal R^2 / Conditional R^2	0.140 / 0.153		

Appendix 2: Varying Intercept of Each State



Appendix 3: Residuals vs Fitted Values of Final Model


Residuals vs Fitted Values
Residuals
Fitted Values