StreetRx

Aarushi Verma, Deekshita Saikia, Mohammad Anas, Tego Chang, and Sydney Donati-Leach

Introduction and Summary

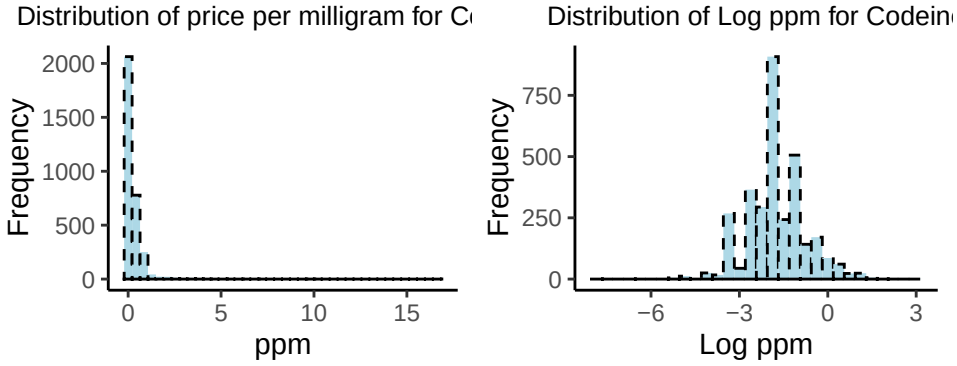
StreetRx (streetrx.com) is a web-based citizen reporting tool enabling real-time collection of street price data on diverted pharmaceutical substances. Users anonymously report prices they paid for prescription drugs on the web. As part of our analysis, we built hierarchical models to investigate how location among other factors may influence the pricing per milligram of a certain drug. According to our analysis, the drug Codeine is found to cheaper in the Michigan, Missouri, Illinois and Texas states. We also found out that purchasing the drug in higher dosages and in huge quantities cad reduce the price paid for Codiene per milligram.

Data

The data used in this analysis pertains to the drug, *Codeine*. We examined our data and observed missing values in columns ppm (price per milligram), mgstr (dosage strength in mg) and source (source of the reported price). We dropped the rows with missing values in ppm and mgstr column. Our original data had 4134 observations for Codeine which reduced to 3125 after missing value removal. Since the source column had high number of missing values and multiple unique values, we categorized the column into *Personal*, *Heard it*, *Internet*, *Not Indicated*. The column mgstr is a discrete variable however, it only contained 3 unique values. We created a new column called dosage to indicate the potency of the drug based on the mgstr values as low, medium or high dosage. In the state column we noted that some states had been incorrectly updated as USA hence, we replaced this value with "Others". We did not consider the form variable in our analysis since it only contained one value - pill/tablet. for the drug Codiene.

EDA

The first step in our EDA was to plot the response variable *ppm* to check whether it follows a normal distribution in order to build a linear regression model. We observed the distribution of *ppm* is highly skewed to the right. We used a log transformation on ppm to address the skewness. The distribution of log ppm was relatively normal and we decided to move ahead with our EDA with log ppm as our response variable.



To explore the data further, we plotted our variables to establish any interesting associations with the response variable. Since all our variables are factor variables, we plotted box plots to identify relationships between our variables.

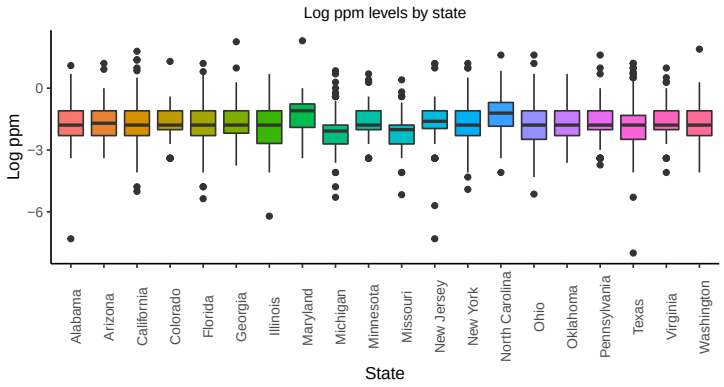
To establish main effects we plotted log ppm against the variables $Source, Bulkpurchase and Dosage$. We did not observe any change in trend in the plots for log ppm vs. Source and Bulk Purchase. We did note a variation in the trend plot for log ppm and dosage indicating an association between the two. However we must also bear in mind that both variables are derived from the price entered by the user for the drug.

After assessing the relationships between the response and explanatory variables, we went on to further explore the interactions between the explanatory variables . The box plots for interaction between Dosage vs Bulk Purchase and Dosage vs. Source showed no change in trend. In particular, we observed some variation in trend of log(ppm) and bulk_purchase when looked at separately for each source. We concluded to explore this interactions further in our model.

We also plotted the response variable against our location variables - State and USA region to establish whether we should include the random intercept for these variables.

We observed a slight trend change between log ppm and USA region. For state, we saw variation across multiple states. Since states are nested within regions, we decided to assess the impact of location on the price of the drug at a more granular level i.e. at the state level. Hence we included state as our hierarchical variable.

The state variable has more than 50 levels, therefore to look at the trend we plotted a subset of these states. We filtered states with more than 50 observations in our data and plotted them against log ppm to see if there was any variation. We did see a variation and thus we decided to control our intercept of the model by state.



We also investigated whether we need to include any random slopes in our model based on the interaction between the main effects and our grouping variable State. We plotted box plots to investigate the associations and concluded that we may need to include random slopes by State for the Source variables and Bulk Purchase variable based on change in trends across states for ppm for these variables.

Model

To build our model, we first built our baseline linear regression which included all our main effects without controlling any of the intercepts or slopes by state. Next we used step wise selection using AIC to generate our final linear model with main effects. We then included some interaction effects which we thought to be significant, or they answered questions with respect to the study. With the help of anova tests we assessed if these interaction were significant to our model. We then incorporated our random intercepts and random slopes and tested them again using annova to arrive on our final model.

Model Building

Our first model included the main effect of every variable. Since all effects are factor variables we did not need to center them to improve our interpretation. Here our response variable is **log ppm** and the predictors are **source, dosage and bulk purchase**. Next we used step wise selection using AIC and BIC to assess which variables should we retain in our model. Based on this, we removed the variable source from our model. Given that source was an interesting variable we also tested it using anova and the p value was insignificant at 95% confidence level.

To ensure our final model is the best fit for our data, we also included the one interaction effect we found interesting during our EDA to the model and used the anova test to conclude whether the interaction between source and bulk purchase had a significant impact on our model or not. However, based on anova the interaction also came out to be insignificant which meant there seemed to be no additional impact of those interactions.

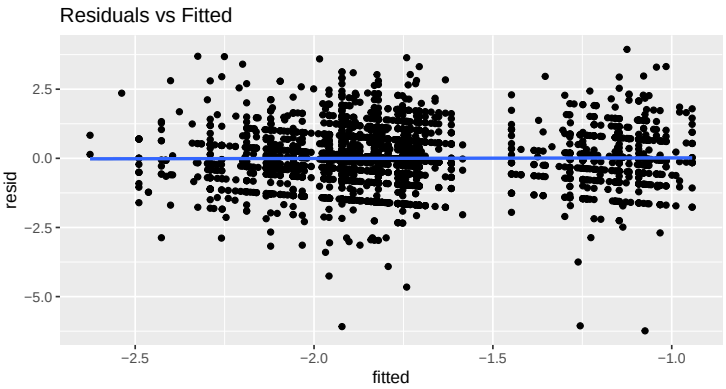
Once we had our linear model, we proceeded to include the random intercepts and slopes that were of interest to us. Based on our EDA we included a varying intercept for the state variable. Further, we also included varying slopes by state for bulk purchase and dosage. In order to analyze whether including varying slopes improved the fit of the model, we used anova to compare each of the varying slope models to our model, which only included the varying intercept. However based on anova, the p value for both the models with varying slopes was insignificant at 95% confidence level indicating that controlling for slopes by state did not improve the fit of our model.

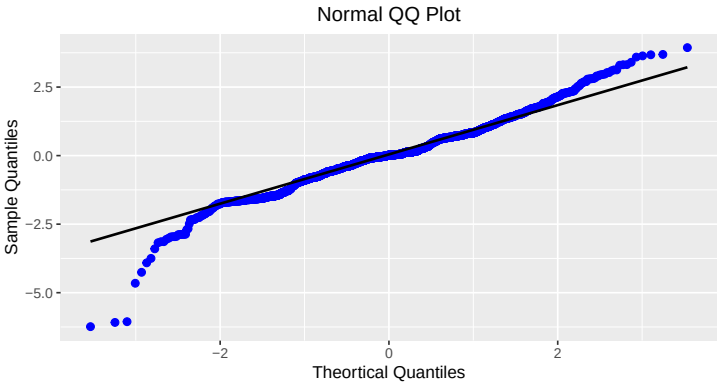
$$y_{i\ state} = (\beta_0 + \gamma_{0\ state}) + (\beta_1) dosage_{1i\ state} + (\beta_1) bulk\_purchase_{1i\ state} + \epsilon_{i\ state};\ i = 1, ..., n_{state};\ state = 1, ..., 59$$

Model Assessment

To assess our final model we checked if the assumptions of Linearity, Normality, Equal variance and Independence were violated. Since all our variables are factor variables, we were unable to verify the linearity assumption.

To check the independence and equal variance assumptions we plotted the residuals against the fitted values. The points seemed randomly distributed with no discernible pattern and the spread of variables seemed constant above and below the line. There did seem to be some points on the x axis that may have violated the equal variance of errors assumptions however, they were only few points and it is safe to say that neither of the above mentioned assumptions were violated. However, this does indicate that there are outliers are present in our data. To check for normality, we plotted the Q-Q plot. For our model we observed that majority of the points lie on the 45 degree line. Both the *Q-Q plot* and the _Residuals vs fitted plot are shown below.





To check if outliers were affecting our model we removed them from our data and ran the model again. However, the standard estimates and their p-values did not change and hence we can conclude that the outliers were not affecting our model.

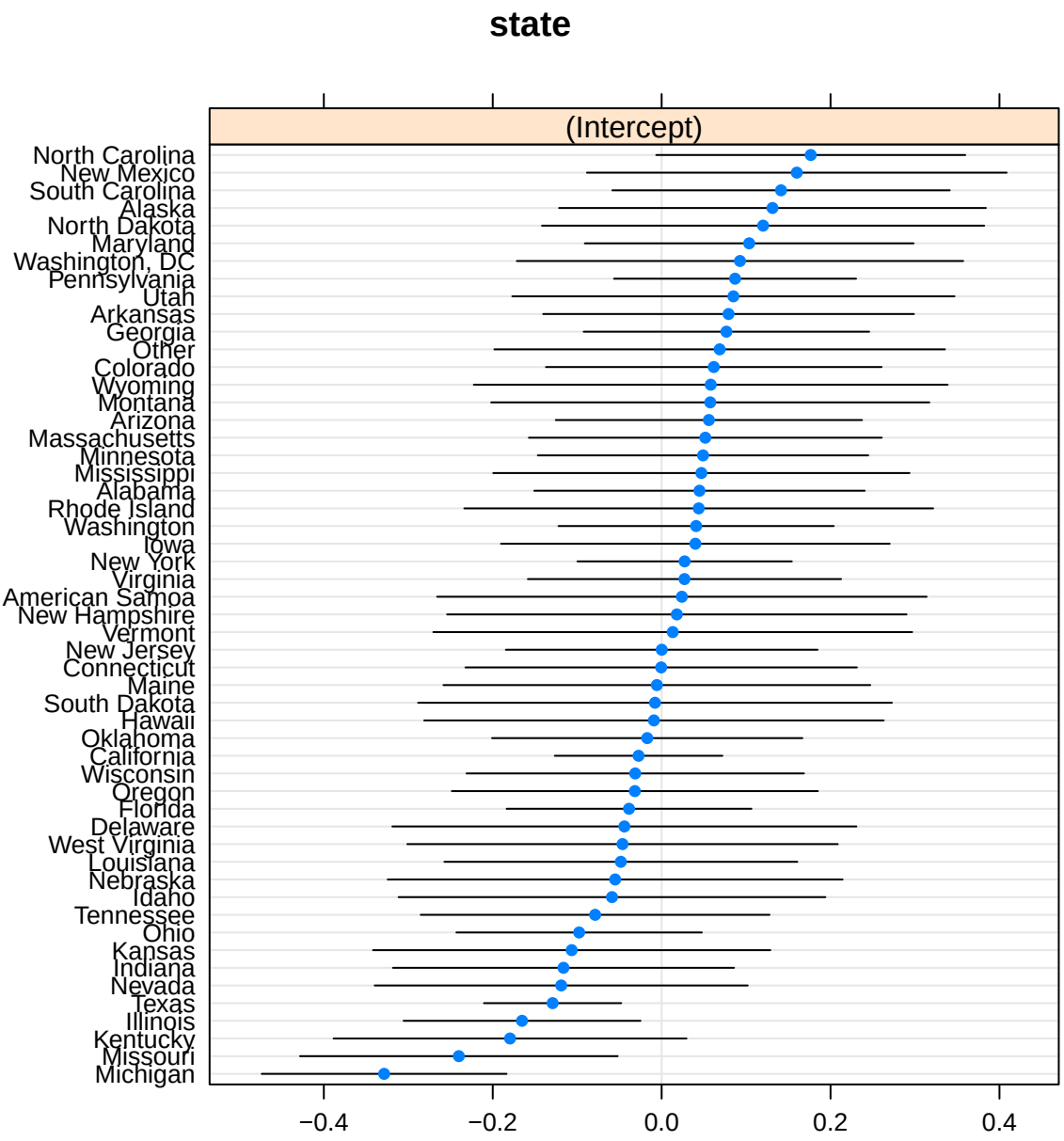
Model Interpretations

We found out that all our explanatory variables had a significant effect on log_ppm. To make the interpretation of the standard estimates simpler we exponentiate them and measure the effect on price. We notice that for low dosage and drug not being purchase in bulk the price was 0.33 USD. If the drug was purchased in high dosage the price per milligram decreased by 65% compared to if it was purchase in low dosage. When purchase in medium dosage the price/mg decreases by 51%. If the drug was purchase in bulk, the price/mg dropped by 87%. We also note that controlling our intercept for state only explains 13.2% of the variation in log(ppm). The summary of our model is shown below.

Table 1: Hierarchical Model Summary

	Dependent variable:
	log(ppm)
dosagehigh	-1.04^{***} $(-1.18, -0.90)$
dosagemedium	-0.67^{***} $(-0.76, -0.58)$
bulk_purchase1 Bulk purchase	-0.14^{***} $(-0.23, -0.05)$
Constant	-1.12^{***} $(-1.21, -1.02)$
Observations	3,215
Log Likelihood	−4,537.65
Akaike Inf. Crit.	9,087.31
Bayesian Inf. Crit.	9,123.76
Note:	*p<0.1; **p<0.05; ***p<0.01

Looking at the dot plot of random effects we notice that the the drug Codeine is found to be cheaper in the states Michigan, Missouri, Illinois and Texas. The confidence interval of the random effect for the other states contains zero, hence, we can say that the price of the drug does not vary significantly for them. The dotplot is shown below.



Limitations and Conclusions

There are a few potential limitations in our model. Firstly, we removed 919 data observations when dealing with missing values. This can be avoided using missing value imputation methods. The second major limitation is that the data used is not reliable as it is crowd sourced and any one can put in any value for the price of the drug.

To conclude, we note that the drug Codeine can be bought for a cheaper price in the Michigan, Missouri, Illinois and Texas states. Purchasing the drug in bulk quantity and at high dosages can also reduce the price of the drug. It was surprising to find out that the source variable did not have a significant affect on the price/mg as we initially hypothesized that prices on the internet to be more expensive than the prices they observed through personal experiences.