Linear Modeling of Diazepam Street Prices in the United States

Xuzhang Li, Yue Li, Derek Tao, Haoxuan Wang, and Yanjiao Yang

2023/10/29

1 Introduction

Prescription opioid division and abuse are major public health issues, and street prices provide an indicator of drug availability, demand, and abuse potential. Using **StreetRx** data, our goal is to investigate the factors related to the price per milligram (mg) of **Diazepam**. **StreetRx** (`streetrx.com`) is a web-based citizen reporting tool enabling real-time collection of street price data on diverted pharmaceutical substances. Based on the principles of crowd-sourcing for public health surveillance, the site allows users to anonymously report prices they paid or heard were paid for diverted prescription drugs. User-generated data offers valuable insights into an otherwise opaque black market, providing a novel dataset for public health surveillance, particularly regarding controlled substances.

The goal of this project is to investigate the factors related to the price per mg of **Diazepam**, accounting for potential grouping by location and exploring heterogeneity in drug pricing across locations. Our objective is to assess the relationships between these factors and the price per mg of the drug.

2 Data Cleaning & EDA

Missing data. We began by examining the missing data in the dataset. Figure 7 provides a summary of the overall missing patterns. Notably, several variables, including `city`, `source`, and `Primary-Reason`, exhibit a significant number of missing values. `Primary-Reason` will not be considered due to its substantial absence of observations before 2016 Q4 and the fact that over 50% of the observations are missing. `Form-temp` is also excluded as all entries are 'pill/tablet'. Furthermore, `city` is not deemed a reliable grouping variable for explaining the heterogeneity of drug prices across locations, primarily due to data entry errors and high missing rates. Since all observations with missing values for drug price (`ppm`) also have missing values for dosage (`mgstr`), we remove all data with missing values for `ppm`. After this removal, there will be no missing values for either `ppm` or `mgstr`.

Variable selection. Based on the frequency of `mgstr`, the majority of values fall within $\{2, 4, 5, 8, 10\}$ mgs. The remaining two levels (`mgstr`=$\{1, 20\}$) have only one observation each, so we remove these two levels and encode `mgstr` as a five-level factor categorical variable. Moreover, to handle the variable `source`, we encode missing values as a new category 'Unknown'. Given that `source` has a high missing rate of approximately 50%, we assess its correlation with ppm using complete data, which yields a low correlation of -0.0081 (see the Appendix for correlation test results). Therefore, we have decided to exclude `source` from the model. Furthermore, we eliminate observations labeled as 'Other/Unknown' in `USA-region`, since they outnumber the `state` entries labeled as 'Other', and all 'Other' entries in `state` correspond to 'Other/Unknown' entries in `USA-region`. Lastly, we exclude `country` as a grouping variable, as all drugs are purchased in the U.S.

Distribution of `ppm`. We observe from Figure 1 that the distribution of `ppm` is heavily right-skewed. To meet the conditional distribution assumption, we choose to apply a log transformation of `ppm` to make the distribution relatively normal and symmetric. This approach is valid since all raw `ppm` are positive and interpretable. Additionally, we identify outliers in log(`ppm`). As the data can be contributed by anyone on the Internet, we remove the nonsense values lying outside the interval [Q1-2.5*IQR,Q3+2.5*IQR], where Q1 and Q3 are the lower and upper quartile of `ppm` (Figure 8 in the Appendix).

Random intercepts. One of our research questions is to understand the pricing heterogeneity of Diazepam by location, which we intend to investigate by incorporating a random intercept based a location covariate. We exclude `city` as a grouping variable due to a lot of missing entries. We evaluate both `state` and `USA-region` as potential grouping variables by examining the

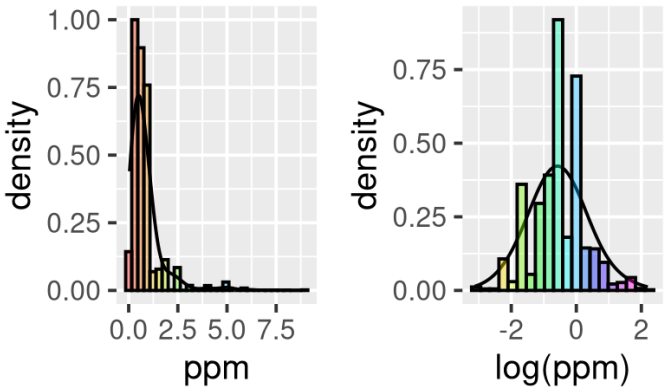

Figure 1: Distribution of `ppm` and log(`ppm`)

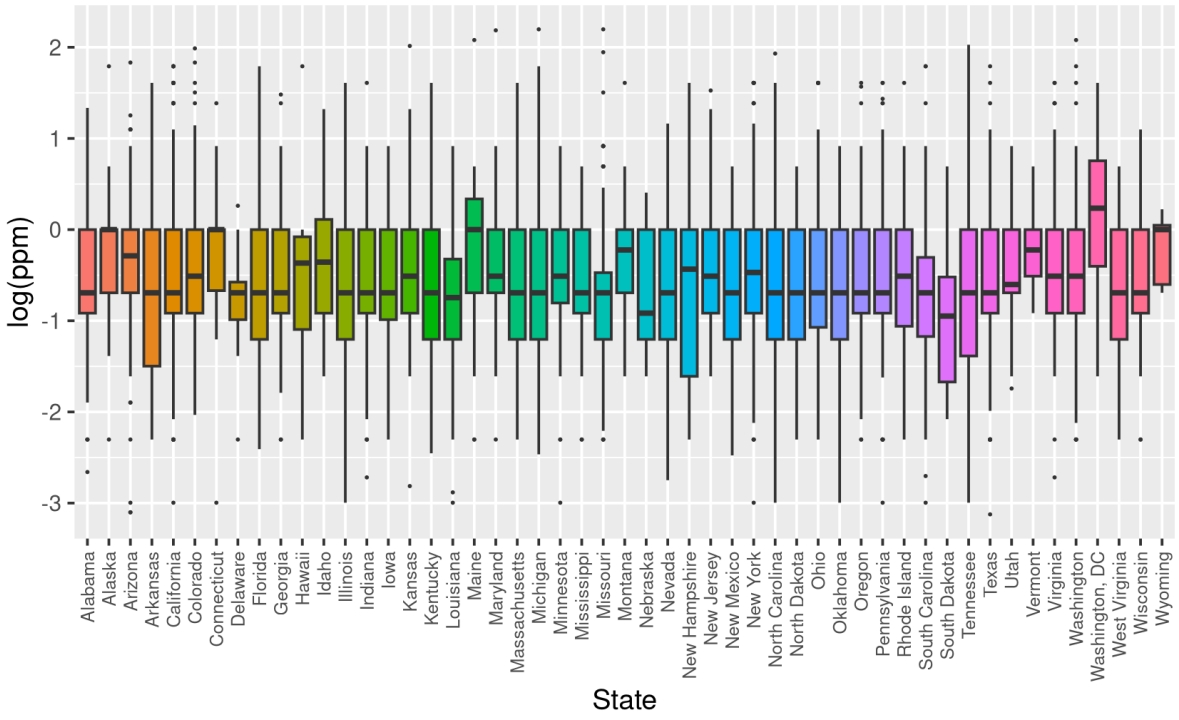

Figure 2: log(ppm) by state

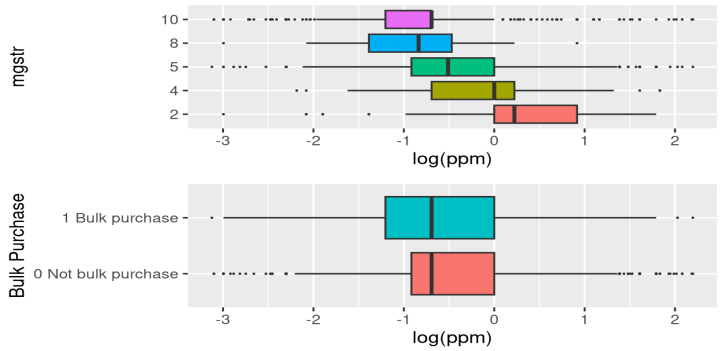

Figure 3: log(ppm) by `mgstr`, `bulk-purchase` and `source`

heterogeneity of log(ppm) among states and regions. From the boxplots provided in Figure 2 and Figure 9 in the appendix, we decide to include `state` as a random intercept because there is substantial variation in log(ppm) across states. We do not consider `USA-region` as the grouping covariate, since there are only 4 regions and it is not enough to properly estimate the between-region variance during modeling.

Fixed effects. Another research question of interest is to identify the covariates associated with ppm of Diazepam. To explore the temporal variation of ppm, we firstly examine the relationship between the year and log(ppm). Notable variations in log(ppm) by year are observed, particularly after 2010. This prompts us to include 'year' as a fixed effect in the model, with the exclusion of pre-2010 observations due to their scarcity (refer to the Appendix Figure 12). However, log(ppm) does not exhibit evident variations by quarter or month, so we exclude them from our model (see Figure 13, 14 in the Appendix). Additionally, both `mgstr` and `bulk-purchase` display differences in log(ppm) across their respective levels (refer to Figure 3), leading us to include them as fixed effects in our model. It's worth mentioning that we plan to set `USA-region` as a fixed effect. By including it, the model will have a hierarchical structure: we basically specify region-specific means, and we could further specify state-specific means following a normal distribution within each region.

3 Model Building and Selection

Through our EDA, we decide to include a random intercept for `state`. Additionally, we incorporate `year`, `mgstr`, and `bulk-purchase` as fixed effects, with `year` being coded as a categorical variable. As stated in our EDA, we have chosen `USA-region` as a fixed effect, and provided our rationale for not building a nested model, which would include both `USA-region` and `state` as random effects. In summary, our model adopts a two-layer hierarchical structure, with `USA-region` at the top level as a fixed effect and `state` at the bottom level as a random intercept. Therefore, in the formula of the `lmer` function in R, we code (1 | USA-region:state) instead of the usual (1 | state).

Secondly, to investigate the necessity of adding random slopes to our model, we explore the inclusion of random slopes for `mgstr`, `bulk-purchase`, and both. Based on the result presented in Table 1, we do not observe any significance in including random slopes in our model (see the Appendix for the code).

In the last step, we apply LASSO and Elastic Net to explore potential model interactions. The advantage of using LASSO is that the ℓ_1 penalty can shrink certain coefficients to exactly zero, effectively performing variable selection. We fit a model with

model	BIC	p-value
m1: log(ppm) ∼ mgstr+bulk-purchase+year+USA-region+(1 | USA-region:state)	10737	-
m2: log(ppm) ∼ mgstr+bulk-purchase+year+USA-region+(1+mgstr | USA-region:state)	10848	0.8849
m3: log(ppm) ∼ mgstr+bulk-purchase+year+USA-region+(1+bulk-purchase | USA-region:state)	10754	0.7304
m4: log(ppm) ∼ mgstr+bulk-purchase+year+USA-region+(1+bulk-purchase+mgstr | USA-region:state)	10897	0.9742

Table 1: Model comparison of models with/without random slopes (p-values are compared to m1)

	Estimate	Std. Error	t value
(Intercept)	-0.3984	0.2286	-1.74
mgstr4	-0.4896	0.0597	-8.19
mgstr5	-0.6799	0.0363	-18.71
mgstr8	-1.1506	0.1433	-8.03
mgstr10	-1.1346	0.0350	-32.43
bulkpurchase1Bulkpurchase	-0.1681	0.0256	-6.57
year2011	0.8669	0.2757	3.14
year2012	0.7324	0.2430	3.01
year2013	0.7652	0.2350	3.26
year2014	0.6555	0.2309	2.84
year2015	0.6034	0.2269	2.66
year2016	0.6466	0.2256	2.87
year2017	0.5845	0.2258	2.59
year2018	0.6168	0.2259	2.73
year2019	0.7318	0.2331	3.14
USAregionNortheast	0.1846	0.0381	4.85
USAregionSouth	0.0363	0.0311	1.17
USAregionWest	0.1595	0.0349	4.57

Table 2: Fixed effects estimates on log scale

all possible interactions among the four fixed effects using LASSO, and then identify which coefficients are shrunk to zero. In detail, we initially construct the design matrix using the four fixed effects, as well as combinations of interactions between them. Subsequently, we conduct cross-validation with 10,000 iterations to determine the optimal tuning parameter λ. Finally, we apply LASSO with this optimal λ and observe the coefficients. The results detailed in the Appendix demonstrate that for each possible interaction, more than 50% of the coefficients are reduced to zero, indicating that interactions are unnecessary in our model.

Based on the three steps above, we determine our final model, with `state` as random intercept and `mgstr`, `bulk-purchase`, `year` and `USA-region` as fixed effects:

$$y_{ijk} = \mu + \sum_{m\in\{4,5,8,10\}} \beta_{1,m} \times \mathbb{I}\{\text{mgstr}_{ijk} = m\} + \beta_2 \times \mathbb{I}\{\text{purchase}_{ijk} = 1\} + \sum_{l\in\{2011,\ldots,2019\}} \beta_{3,l} \times \mathbb{I}\{\text{year}_{ijk} = l\} \\ + \beta_{4,k} + \alpha_{j,k} + \varepsilon_{ijk}, \quad \alpha_{j,k} \overset{iid}{\sim} \mathcal{N}(0,\tau^2), \quad \varepsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0,\sigma^2), \quad i = 1,\ldots,n_{jk},\ j = 1,\ldots,n_k,\ k = 1,2,3, \tag{1}$$

where y_{ijk} is the log(`ppm`) for purchase i in state j in region k, `mgstr` is the dosage strength in mg, μ is the grand mean, n_{jk} is the number of individuals within state j of region k, n_j is the number of states within region k, $\alpha_{j,k}$ is the random effect for state j in region k, $k = 1$ represents `USA-region`=Northeast, $k = 2$ represents `USA-region`=South, $k = 3$ represents `USA-region`=West, ε_{ijk} is the residual unexplained by covariates, τ^2 is the within-region variability, and σ^2 is the variability of purchases within state j of region k. R automatically uses `mgstr`=2, no `bulk-purchase`, `year`=2010, and `USA-region`=Midwest as the reference level.

4 Model Diagnostics and Limitations

There are a few limitations to our model. The within-region variance is 0.0011, which is much smaller than the within-state variance, 0.5039 (Table 3). This means there is still much within-state variance that our model is unable to explain. Additionally, there are issues with the normality assumptions and influential groups explained below.

Residual analysis. Through examining the residuals of our model, we can determine how well our model assumptions hold. In Figures 4 and 5, we have included the residual versus fitted plot and the residual distributions, which we can use to evaluate our model assumptions. Based on the residual versus fitted plot, we can see that there is a clear negative linear, non-random trend in the residuals, which indicates a violation of the independence assumption. We can also see that there are slight differences in the variance of the residuals across the fitted values, which indicates a potential violation of the homoscedasticity assumption, though this violation is not too extreme in this case. In Figure 5, the residual distribution shows deviation from normal behavior in the tails of the distribution, indicating a likely violation of the normality assumption.

Influential analysis. We determine if there are influential groups that might be affecting our model behavior. We have included

Source of variation	Estimate
State	0.0011
Residual	0.5039

Table 3: Variance estimate of the model

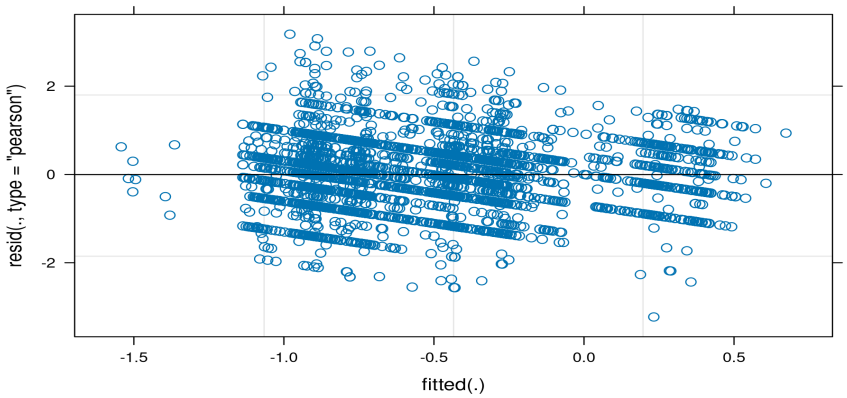

Figure 4: Residual versus fitted

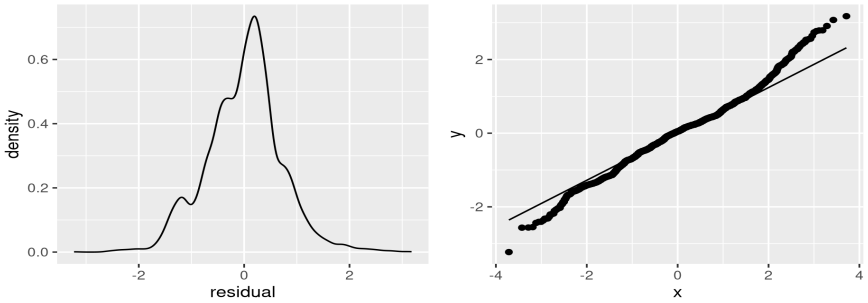

Figure 5: Residual distributions

the DFBETAS of each model parameter for each state, as well as Cook's distance by state, to check for influential states in our model.

For DFBETAS (Figure 15 in the appendix), we use a recommended cutoff of $2/\sqrt{n} = 0.28$. Based on this metric, we observe that 20 of the states exceed this cutoff for at least one of our model's parameters. However, because our model is complex, we put less consideration into DFBETAS and more into Cook's Distance, since we consider exceeding the cutoff for a single parameter in DFBETAS to be less significant than exceeding the cutoff for Cook's Distance.

For Cook's Distance (Figure 16 in the appendix), we use a recommended cutoff of $4/n = 0.0784$. The one state that exceeds this threshold is New York, which also exceeds the DFBETAS threshold and can be considered an influential point. We also notice that the states with highest Cook's Distance are also included in the states deemed influential by DFBETAS.

Something else of note is that the states with highest Cook's Distance (ie. California, New York, Pennsylvania) are also the states with some of the highest sample sizes in the data. It is possible that these high sample sizes are partially causing the high influence in these states, since removing a larger amount of data from the model would likely result in more significant changes to model parameters.

Data Limitations. We also would like to note some limitations of the data itself, which may have impacted the results of our model. One such limitation relates to the `source` variable, which is a feature that we initially did deem to be important from both a modeling and real-world standpoint. However, we ultimately had to remove it entirely because of the high rate of missing data. Thus, our model lacks explanation of how information source impacts ppm, and more importantly, how the information source can possibly be an indicator of bias in ppm (ie. perhaps there is much higher variation in reported ppm from the internet than from in-person purchases). Another limitation within the data is that the `bulk-purchase` variable, which is an indicator of the volume of a drug purchase, has only two levels with a somewhat arbitrary cutoff between bulk and non-bulk. From a real-world standpoint, the volume of a drug purchase would seemingly be highly associated with ppm, so in order to improve the precision of our analysis, it would have been beneficial to have more numerous and specific levels to the `bulk-purchase` feature.

5 Interpretation of Results

Fixed effects. In Table 2, each estimate in our model is significant with 0.05 significance level. Therefore, we have identified a relationship between `mgstr`, `bulk-purchase`, `year`, `USA-region` and the price per mg of Diazepam. As we take the log of our response `ppm`, we need to exponentiate the estimates to interpret the effect of each variable on `ppm`. Below are the interpretations:

- Grand mean: Our estimated grand mean for `ppm` is 0.6714 ($= e^{-0.3984}$). This is the average price of Diazepam across all states, when dosage strength 2 mg of units of Diazepam purchased (`mgstr`), without bulk purchase, purchased in 2010, and purchased in Midwest USA.
- `mgstr`: On average, holding other variables constant, we see a total increase in dosage from 2 mg to 10 mg results in a 47.53% ($= e^{-1.1346}/e^{-0.4896} - 1$) decrease in price per mg of Diazepam. In particular, compared to a dosage of 2 mg,

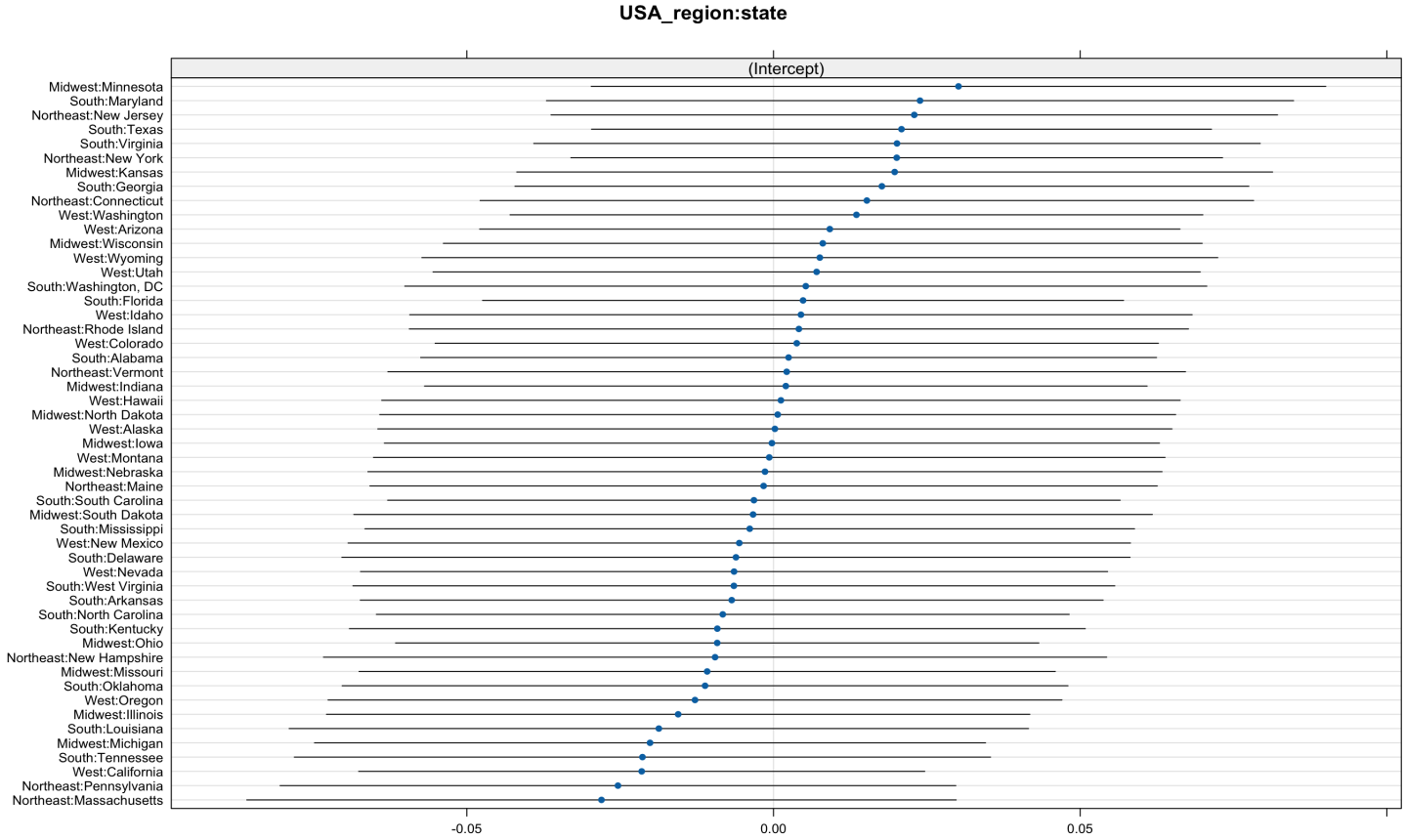

Figure 6: 95% confidence intervals for random intercept by state

increasing the dosage from 5 mg to 8 mg results in the greatest decrease, 37.54%, in price per mg of Diazepam. Increasing the dosage from 8 mg to 10 mg results in a slight 1.61% increase in price per mg of Diazepam.

- `bulk-purchase`: On average, holding other variables constant, bulk purchases (purchases of more than 10 units of Diazepam at once) are 15.47% ($= 1 - e^{-0.1681}$) cheaper, in terms of price per mg, than non bulk purchases (purchase less than 10 units of Diazepam at once).
- `year`: On average, holding other variables constant, price per mg of Diazepam decreases 23.16% (totally) from 2010 to 2015, and increases 13.7% (totally) from 2015 to 2019.
- `USA-region`: On average, holding other variables constant, compared to purchasing Diazepam in Midwest USA, purchasing it in Northeast USA leads to the largest increase in price per mg, 1.2027, while purchasing it in Southern USA leads to the smallest increase in price per mg, 1.037.

Random effects. Figure 6 illustrates the sorted random state intercepts with 95% confidence intervals. Based on this plot, we can observe the average impacts of each state on ppm of Diazepam in our model. Specifically, we see that Minnesota, Maryland, and New Jersey have the highest average ppm of Diazepam, while Massachusetts, Pennsylvania, and California have the lowest average ppm of Diazepam. We see that the distance from the state to the regions means (the 0 intercept) roughly resembles a normal distribution with mean 0 and standard deviation 0.01-0.015, representing a small effect from state level to price per milligram. Also, all of the 95 percent confidence intervals for the state random intercepts include the grand mean of 0, which implies that between-state variance is low in our model. Table 3 summarizes the across-state and within-state variance of our model. As mentioned in the model limitations, there is still a lot of within-state variance that our model is unable to explain.

General conclusion. We now give a general view of our results. We first established a grand mean price per mg of Diazepam at the baseline of our model parameters. We then observed that increasing the dosage of Diazepam mostly results in a decrease in the price per mg, except in the two highest dosage amounts. We also concluded that bulk purchases of Diazepam will lower the price per mg. As for trends of price per mg over the years, we observed that there isn't a consistent positive or negative trend year-to-year from 2010 to 2019, though we could say that there has been an overall decrease in price per mg from 2010 to 2019. We were also able to determine how each U.S. region impacts ppm, as well as how each state within each region impacts ppm.

In terms of how Diazepam price per mg differs across or within states, we concluded that different states do not have very noticeable differences in price per mg, while there may be significant differences in price per mg within a single state. We also identified some states that have high influence on our model results, and acknowledged that their high influence may stem from the fact that these states have large populations and thus a higher individual response rate within the data.

6 Appendix

6.1 Code

```
### Part 1: data cleaning ###
# load data
load("streetrx.RData")
# filter data for only diazepam
streetrx <- streetrx %>%
  filter(api_temp == "diazepam")
  # code missing data
streetrx <- data.frame(apply(streetrx, 2, function(x) gsub("^$|^␣$", NA, x)))
# change dtypes for each column
streetrx <- streetrx %>%
  mutate(ppm = as.numeric(ppm),
         yq_pdate = as.numeric(yq_pdate),
         price_date = mdy(price_date),
         city = as.factor(city),
         state = as.factor(state),
         country = as.factor(country),
         USA_region = as.factor(USA_region),
         source = as.factor(source),
         form_temp = as.factor(form_temp),
         mgstr = as.numeric(mgstr),
         bulk_purchase = as.factor(bulk_purchase),
         Primary_Reason = as.factor(Primary_Reason))
# add year
streetrx <- streetrx %>%
  mutate(year = year(price_date)) %>%
  filter(year>2009)

# delete observations with missing ppm data
streetrx <- streetrx %>%
  filter(!is.na(ppm))

# delete USA_region=Unknown, since # USA_region missing > # state other
streetrx <- streetrx %>%
  filter(USA_region != "Other/Unknown")

# delete levels of mgstr that are not 2,4,5,8,10, since there are only 2 of them
streetrx <- streetrx %>%
  filter(mgstr %in% c(2,4,5,8,10)) %>%
  mutate(mgstr = as.factor(mgstr))

# replace NA's in source with "Unknown"
streetrx <- streetrx %>%
  mutate(source = as.character(source)) %>%
  mutate(source = if_else(is.na(source), "Unknown", source))

# combine all website sources as being 'Internet Source'
streetrx <- streetrx %>%
  mutate(source = if_else(str_detect(source, "http://"), "Internet", source)) %>%
  mutate(source = if_else(str_detect(source, ".com$"), "Internet", source)) %>%
  mutate(source = if_else(source == "Internet␣Pharmacy", "Internet", source)) %>%
  filter(source != "fuck␣u") %>%
  filter(source != "Eronhkkjhhhhjj")

# Removing Outliers
# Compute the log-transformed value of ppm
streetrx$log_ppm <- log(streetrx$ppm)

# Compute Q1, Q3, and IQR for log(ppm)
Q1_R <- quantile(streetrx$log_ppm, 0.25, na.rm = TRUE)
Q3_R <- quantile(streetrx$log_ppm, 0.75, na.rm = TRUE)
IQR_R <- Q3_R - Q1_R

# Define the bounds for outliers using a multiplier of 3
lower_bound_R <- Q1_R - 2.5 * IQR_R
upper_bound_R <- Q3_R + 2.5 * IQR_R

# Count the number of observations outside the bounds
outliers_below_R <- sum(streetrx$log_ppm < lower_bound_R, na.rm = TRUE)
```

```r
outliers_above_R <- sum(streetrx$log_ppm > upper_bound_R, na.rm = TRUE)
total_outliers_R <- outliers_below_R + outliers_above_R

# Print the number of outliers
cat("Number of outliers below the lower bound:", outliers_below_R, "\n")
cat("Number of outliers above the upper bound:", outliers_above_R, "\n")
cat("Total number of outliers:", total_outliers_R, "\n")
## Number of outliers below the lower bound: 27
## Number of outliers above the upper bound: 34
## Total number of outliers: 61

streetrx_outrm <- streetrx %>%
  filter(log_ppm >= lower_bound_R & log_ppm <= upper_bound_R)

#Remove Outliers
streetrx <- streetrx_outrm

# save the cleaned & transformed data
saveRDS(streetrx, "streetrx_cleaned.RData")
### Part 2: EDA ###

ppm_hist1 <- streetrx %>%
  ggplot(aes(x = ppm, y = ..density..)) +
  geom_histogram(alpha = 0.4, fill = rainbow(30), bins = 30, color = "black") +
  geom_density(color = "black", adjust = 5)

## the distribution of log(ppm) looks approximately normal
ppm_hist2 <- streetrx %>%
  ggplot(aes(x = log(ppm), y = ..density..)) +
  geom_density(color = "black", adjust = 5) +
  geom_histogram(alpha = 0.4, fill = rainbow(20), bins = 20, color = "black")

ppm_hist1 + ppm_hist2

ggplot(streetrx, aes(x = state, y = log(ppm))) +
  geom_boxplot(aes(fill = factor(state)), outlier.size = 0.1) +
  labs(x = "State") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1, size =

ggplot(streetrx, aes(x = USA_region, y = log(ppm))) +
  geom_boxplot(aes(fill = factor(USA_region)), outlier.size = 0.1) +
  labs(x = "Region") +
  theme(legend.position = "none", axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))

ggplot(streetrx, aes(x=log_ppm)) +
  geom_histogram(aes(y=..density..), bins=50, fill="lightblue", color="black", alpha=0.7) +
  geom_density(color="red") +
  geom_vline(aes(xintercept=lower_bound_R), linetype="dashed", color="green", size=1) +
  geom_vline(aes(xintercept=upper_bound_R), linetype="dashed", color="blue", size=1) +
  labs(title="Distribution of log(ppm) with Outlier Bounds",
       x="log(ppm)", y="Density") +
  theme_minimal()

streetrx %>% filter(yq_pdate >= 20000) %>%
  mutate(year = yq_pdate %/% 10) %>%
  ggplot() +
  geom_boxplot(aes(x = year, y = log(ppm), group = year))

streetrx %>%
  mutate(quarter = yq_pdate %% 10) %>%
  ggplot(aes(fill = factor(quarter))) +
  geom_boxplot(aes(x = quarter, y = log(ppm), group = quarter), outlier.size = 0.1) +
  theme(legend.position = "none") +
  labs(x = "Quarter")

mgstr_ppm <- ggplot(streetrx, aes(x = factor(mgstr), y = log(ppm))) +
  geom_boxplot(aes(fill = factor(mgstr)), outlier.size = 0.1) +
  labs(x = "mgstr") +
  theme(legend.position = "none") +
  coord_flip()
bp_ppm <- ggplot(streetrx, aes(x = bulk_purchase, y = log(ppm))) +
  geom_boxplot(aes(fill = bulk_purchase), outlier.size = 0.1) +
```

```
  labs(x = "Bulk Purchase") +
  theme(legend.position = "none") +
  coord_flip()
mgstr_ppm / bp_ppm

ggplot(streetrx, aes(x = source, y = log(ppm))) +
  geom_boxplot(aes(fill = source), outlier.size = 0.1) +
  labs(x = "Source") +
  theme(legend.position = "none")

# source: small cor
streetrx_source = streetrx %>% filter(source != "Unknown") %>%
  mutate(source=if_else(source=="Personal", 1,
                        if_else(source=="Internet", 2, 3)))
cor.test(streetrx_source$source, streetrx_source$ppm)
##
## Pearsons product-moment correlation
##
## data: streetrx_source$source and streetrx_source$ppm
## t = 0.33039, df = 2774, p-value = 0.7411
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.03093685 0.04346532
## sample estimates:
## cor
## 0.006272914
cor(streetrx_source$source, streetrx_source$ppm)
## [1] 0.006272914

### Part 3: Model selection & comparison ###
# convert year to categorical variables
streetrx <- streetrx %>%
  mutate(year=as.factor(year))
# get log_ppm
streetrx$log_ppm <- log(streetrx$ppm)
mod1 <- lmer(log(ppm) ~ mgstr+bulk_purchase+year+USA_region+(1|USA_region:state), data = streetrx, REM
mod2 <- lmer(log(ppm) ~ mgstr+bulk_purchase+year+USA_region+(1+mgstr|USA_region:state), data = streetrx
anova(mod2, mod1)
mod3 <- lmer(log(ppm) ~ mgstr+bulk_purchase+year+USA_region+(1+bulk_purchase|USA_region:state), data =
anova(mod3, mod1)
mod4 <- lmer(log(ppm) ~ mgstr+bulk_purchase+year+USA_region+(1+bulk_purchase+mgstr|USA_region:state),
anova(mod4, mod1)

# Extract relevant predictors and the original ppm
streetrx$USA_region = factor(streetrx$USA_region) # to remove "Unknown" level from USA_region
levels(streetrx$USA_region)
data <-  streetrx %>% select(year, USA_region, mgstr, bulk_purchase, ppm, state)

formula = as.formula("log(ppm) ~ year * mgstr * bulk_purchase * USA_region")
attr(terms.formula(formula), "term.labels")

data_dummies = model.matrix(formula, data)

response = as.matrix(log(data$ppm))

cv_fit = glmnet::cv.glmnet(x=data_dummies, y=response, maxit=10000)

# Display the optimal lambda and corresponding coefficients
optimal_lambda = cv_fit$lambda.min
lasso_coefficients = coef(cv_fit, s=optimal_lambda)
# print(lasso_coefficients)
# print(optimal_lambda)
lasso.mod<-glmnet::glmnet(data_dummies, response, lambda=optimal_lambda)
coef(lasso.mod)
## 401 x 1 sparse Matrix of class "dgCMatrix"
## s0
## (Intercept) 0.1311167886
## (Intercept) .
## year2011 .
## year2012 .
## year2013 0.1725128456
## year2014 .
```

```
## year2015 .
## year2016 .
## year2017 -0.0149608795
## year2018 .
## year2019 0.0143075203
## mgstr4 -0.2682442813
## mgstr5 -0.5149699779
## mgstr8 -0.6975813006
## mgstr10 -0.9800087896
## bulk_purchase1 Bulk purchase -0.1065961144
## USA_regionNortheast 0.1036617887
## USA_regionSouth .
## USA_regionWest 0.0706992831
## year2011:mgstr4 .
## year2012:mgstr4 .
## year2013:mgstr4 0.4735722676
## year2014:mgstr4 .
## year2015:mgstr4 .
## year2016:mgstr4 .
## year2017:mgstr4 .
## year2018:mgstr4 -0.0908537807
## year2019:mgstr4 .
## year2011:mgstr5 .
## year2012:mgstr5 .
## year2013:mgstr5 .
## year2014:mgstr5 .
## year2015:mgstr5 .
## year2016:mgstr5 .
## year2017:mgstr5 -0.0511990394
## year2018:mgstr5 -0.0405669584
## year2019:mgstr5 .
## year2011:mgstr8 .
## year2012:mgstr8 .
## year2013:mgstr8 -0.2110041798
## year2014:mgstr8 .
## year2015:mgstr8 -0.0698348307
## year2016:mgstr8 .
## year2017:mgstr8 -0.2332318558
## year2018:mgstr8 .
## year2019:mgstr8 .
## year2011:mgstr10 .
## year2012:mgstr10 .
## year2013:mgstr10 -0.1457298379
## year2014:mgstr10 -0.0247469597
## year2015:mgstr10 .
## year2016:mgstr10 .
## year2017:mgstr10 .
## year2018:mgstr10 .
## year2019:mgstr10 0.0198586923
## year2011:bulk_purchase1 Bulk purchase .
## year2012:bulk_purchase1 Bulk purchase .
## year2013:bulk_purchase1 Bulk purchase .
## year2014:bulk_purchase1 Bulk purchase .
## year2015:bulk_purchase1 Bulk purchase .
## year2016:bulk_purchase1 Bulk purchase -0.0289165136
## year2017:bulk_purchase1 Bulk purchase .
## year2018:bulk_purchase1 Bulk purchase .
## year2019:bulk_purchase1 Bulk purchase .
## mgstr4:bulk_purchase1 Bulk purchase .
## mgstr5:bulk_purchase1 Bulk purchase .
## mgstr8:bulk_purchase1 Bulk purchase .
## mgstr10:bulk_purchase1 Bulk purchase .
## year2011:USA_regionNortheast 0.7812593361
## year2012:USA_regionNortheast 0.0812303475
## year2013:USA_regionNortheast 0.0122761032
## year2014:USA_regionNortheast .
## year2015:USA_regionNortheast .
## year2016:USA_regionNortheast .
## year2017:USA_regionNortheast .
## year2018:USA_regionNortheast .
## year2019:USA_regionNortheast .
## year2011:USA_regionSouth .
```

```
## year2012:USA_regionSouth .
## year2013:USA_regionSouth .
## year2014:USA_regionSouth .
## year2015:USA_regionSouth .
## year2016:USA_regionSouth .
## year2017:USA_regionSouth .
## year2018:USA_regionSouth .
## year2019:USA_regionSouth .
## year2011:USA_regionWest .
## year2012:USA_regionWest .
## year2013:USA_regionWest .
## year2014:USA_regionWest .
## year2015:USA_regionWest .
## year2016:USA_regionWest 0.0377603595
## year2017:USA_regionWest .
## year2018:USA_regionWest .
## year2019:USA_regionWest .
## mgstr4:USA_regionNortheast -0.0530694195
## mgstr5:USA_regionNortheast .
## mgstr8:USA_regionNortheast .
## mgstr10:USA_regionNortheast 0.0086308413
## mgstr4:USA_regionSouth .
## mgstr5:USA_regionSouth .
## mgstr8:USA_regionSouth .
## mgstr10:USA_regionSouth -0.0138694626
## mgstr4:USA_regionWest .
## mgstr5:USA_regionWest .
## mgstr8:USA_regionWest -0.1017483750
## mgstr10:USA_regionWest .
## bulk_purchase1 Bulk purchase:USA_regionNortheast .
## bulk_purchase1 Bulk purchase:USA_regionSouth .
## bulk_purchase1 Bulk purchase:USA_regionWest .
## year2011:mgstr4:bulk_purchase1 Bulk purchase .
## year2012:mgstr4:bulk_purchase1 Bulk purchase .
## year2013:mgstr4:bulk_purchase1 Bulk purchase .
## year2014:mgstr4:bulk_purchase1 Bulk purchase .
## year2015:mgstr4:bulk_purchase1 Bulk purchase .
## year2016:mgstr4:bulk_purchase1 Bulk purchase .
## year2017:mgstr4:bulk_purchase1 Bulk purchase .
## year2018:mgstr4:bulk_purchase1 Bulk purchase .
## year2019:mgstr4:bulk_purchase1 Bulk purchase .
## year2011:mgstr5:bulk_purchase1 Bulk purchase .
## year2012:mgstr5:bulk_purchase1 Bulk purchase 1.3495743588
## year2013:mgstr5:bulk_purchase1 Bulk purchase .
## year2014:mgstr5:bulk_purchase1 Bulk purchase .
## year2015:mgstr5:bulk_purchase1 Bulk purchase .
## year2016:mgstr5:bulk_purchase1 Bulk purchase .
## year2017:mgstr5:bulk_purchase1 Bulk purchase .
## year2018:mgstr5:bulk_purchase1 Bulk purchase .
## year2019:mgstr5:bulk_purchase1 Bulk purchase -0.0920781790
## year2011:mgstr8:bulk_purchase1 Bulk purchase .
## year2012:mgstr8:bulk_purchase1 Bulk purchase .
## year2013:mgstr8:bulk_purchase1 Bulk purchase .
## year2014:mgstr8:bulk_purchase1 Bulk purchase .
## year2015:mgstr8:bulk_purchase1 Bulk purchase -1.3575541465
## year2016:mgstr8:bulk_purchase1 Bulk purchase .
## year2017:mgstr8:bulk_purchase1 Bulk purchase .
## year2018:mgstr8:bulk_purchase1 Bulk purchase 0.0265571744
## year2019:mgstr8:bulk_purchase1 Bulk purchase .
## year2011:mgstr10:bulk_purchase1 Bulk purchase .
## year2012:mgstr10:bulk_purchase1 Bulk purchase .
## year2013:mgstr10:bulk_purchase1 Bulk purchase -0.1737042333
## year2014:mgstr10:bulk_purchase1 Bulk purchase .
## year2015:mgstr10:bulk_purchase1 Bulk purchase -0.1407279005
## year2016:mgstr10:bulk_purchase1 Bulk purchase -0.0112942116
## year2017:mgstr10:bulk_purchase1 Bulk purchase .
## year2018:mgstr10:bulk_purchase1 Bulk purchase .
## year2019:mgstr10:bulk_purchase1 Bulk purchase .
## year2011:mgstr4:USA_regionNortheast .
## year2012:mgstr4:USA_regionNortheast .
## year2013:mgstr4:USA_regionNortheast .
## year2014:mgstr4:USA_regionNortheast .
```

```
## year2015:mgstr4:USA_regionNortheast .
## year2016:mgstr4:USA_regionNortheast .
## year2017:mgstr4:USA_regionNortheast .
## year2018:mgstr4:USA_regionNortheast -0.0704544693
## year2019:mgstr4:USA_regionNortheast .
## year2011:mgstr5:USA_regionNortheast 0.1349095794
## year2012:mgstr5:USA_regionNortheast .
## year2013:mgstr5:USA_regionNortheast .
## year2014:mgstr5:USA_regionNortheast .
## year2015:mgstr5:USA_regionNortheast .
## year2016:mgstr5:USA_regionNortheast .
## year2017:mgstr5:USA_regionNortheast .
## year2018:mgstr5:USA_regionNortheast .
## year2019:mgstr5:USA_regionNortheast .
## year2011:mgstr8:USA_regionNortheast .
## year2012:mgstr8:USA_regionNortheast .
## year2013:mgstr8:USA_regionNortheast .
## year2014:mgstr8:USA_regionNortheast .
## year2015:mgstr8:USA_regionNortheast 0.6986114769
## year2016:mgstr8:USA_regionNortheast .
## year2017:mgstr8:USA_regionNortheast .
## year2018:mgstr8:USA_regionNortheast .
## year2019:mgstr8:USA_regionNortheast .
## year2011:mgstr10:USA_regionNortheast .
## year2012:mgstr10:USA_regionNortheast 0.0112975845
## year2013:mgstr10:USA_regionNortheast .
## year2014:mgstr10:USA_regionNortheast .
## year2015:mgstr10:USA_regionNortheast .
## year2016:mgstr10:USA_regionNortheast .
## year2017:mgstr10:USA_regionNortheast 0.0220729001
## year2018:mgstr10:USA_regionNortheast .
## year2019:mgstr10:USA_regionNortheast .
## year2011:mgstr4:USA_regionSouth .
## year2012:mgstr4:USA_regionSouth .
## year2013:mgstr4:USA_regionSouth .
## year2014:mgstr4:USA_regionSouth .
## year2015:mgstr4:USA_regionSouth .
## year2016:mgstr4:USA_regionSouth .
## year2017:mgstr4:USA_regionSouth .
## year2018:mgstr4:USA_regionSouth .
## year2019:mgstr4:USA_regionSouth .
## year2011:mgstr5:USA_regionSouth .
## year2012:mgstr5:USA_regionSouth .
## year2013:mgstr5:USA_regionSouth .
## year2014:mgstr5:USA_regionSouth .
## year2015:mgstr5:USA_regionSouth .
## year2016:mgstr5:USA_regionSouth .
## year2017:mgstr5:USA_regionSouth .
## year2018:mgstr5:USA_regionSouth .
## year2019:mgstr5:USA_regionSouth 0.0657516733
## year2011:mgstr8:USA_regionSouth .
## year2012:mgstr8:USA_regionSouth .
## year2013:mgstr8:USA_regionSouth .
## year2014:mgstr8:USA_regionSouth .
## year2015:mgstr8:USA_regionSouth -0.0015838066
## year2016:mgstr8:USA_regionSouth .
## year2017:mgstr8:USA_regionSouth .
## year2018:mgstr8:USA_regionSouth .
## year2019:mgstr8:USA_regionSouth .
## year2011:mgstr10:USA_regionSouth .
## year2012:mgstr10:USA_regionSouth .
## year2013:mgstr10:USA_regionSouth .
## year2014:mgstr10:USA_regionSouth .
## year2015:mgstr10:USA_regionSouth .
## year2016:mgstr10:USA_regionSouth .
## year2017:mgstr10:USA_regionSouth -0.0275017418
## year2018:mgstr10:USA_regionSouth -0.0262917736
## year2019:mgstr10:USA_regionSouth .
## year2011:mgstr4:USA_regionWest .
## year2012:mgstr4:USA_regionWest .
## year2013:mgstr4:USA_regionWest .
## year2014:mgstr4:USA_regionWest .
```

```
## year2015:mgstr4:USA_regionWest -0.1360814082
## year2016:mgstr4:USA_regionWest .
## year2017:mgstr4:USA_regionWest 0.0776196612
## year2018:mgstr4:USA_regionWest .
## year2019:mgstr4:USA_regionWest .
## year2011:mgstr5:USA_regionWest .
## year2012:mgstr5:USA_regionWest 0.0794332447
## year2013:mgstr5:USA_regionWest .
## year2014:mgstr5:USA_regionWest .
## year2015:mgstr5:USA_regionWest -0.0012295625
## year2016:mgstr5:USA_regionWest .
## year2017:mgstr5:USA_regionWest .
## year2018:mgstr5:USA_regionWest .
## year2019:mgstr5:USA_regionWest .
## year2011:mgstr8:USA_regionWest .
## year2012:mgstr8:USA_regionWest .
## year2013:mgstr8:USA_regionWest .
## year2014:mgstr8:USA_regionWest .
## year2015:mgstr8:USA_regionWest -0.6617520533
## year2016:mgstr8:USA_regionWest .
## year2017:mgstr8:USA_regionWest .
## year2018:mgstr8:USA_regionWest .
## year2019:mgstr8:USA_regionWest .
## year2011:mgstr10:USA_regionWest -0.3685861282
## year2012:mgstr10:USA_regionWest -0.0348575317
## year2013:mgstr10:USA_regionWest .
## year2014:mgstr10:USA_regionWest .
## year2015:mgstr10:USA_regionWest .
## year2016:mgstr10:USA_regionWest .
## year2017:mgstr10:USA_regionWest .
## year2018:mgstr10:USA_regionWest 0.0172623168
## year2019:mgstr10:USA_regionWest 0.0294019276
## year2011:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2012:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2013:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2014:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2015:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2016:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2017:bulk_purchase1 Bulk purchase:USA_regionNortheast 0.0416193783
## year2018:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2019:bulk_purchase1 Bulk purchase:USA_regionNortheast -0.1132715019
## year2011:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2012:bulk_purchase1 Bulk purchase:USA_regionSouth 0.2049632129
## year2013:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2014:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2015:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2016:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2017:bulk_purchase1 Bulk purchase:USA_regionSouth -0.0319250274
## year2018:bulk_purchase1 Bulk purchase:USA_regionSouth 0.0061855912
## year2019:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2011:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2012:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2013:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2014:bulk_purchase1 Bulk purchase:USA_regionWest -0.0963956456
## year2015:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2016:bulk_purchase1 Bulk purchase:USA_regionWest -0.0722492139
## year2017:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2018:bulk_purchase1 Bulk purchase:USA_regionWest 0.1499787097
## year2019:bulk_purchase1 Bulk purchase:USA_regionWest .
## mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast -0.1000084607
## mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth -0.1241492463
## mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2011:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
```

```
## year2012:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2013:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2014:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2015:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2016:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2017:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2018:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2019:mgstr4:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2011:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2012:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2013:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2014:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2015:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2016:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2017:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2018:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast -0.1534075987
## year2019:mgstr5:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2011:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2012:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2013:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2014:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2015:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2016:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2017:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2018:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2019:mgstr8:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2011:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2012:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast 0.0003233928
## year2013:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2014:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2015:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2016:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2017:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2018:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2019:mgstr10:bulk_purchase1 Bulk purchase:USA_regionNortheast .
## year2011:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2012:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2013:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2014:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2015:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2016:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2017:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2018:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth 0.0246792278
## year2019:mgstr4:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2011:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2012:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2013:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2014:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2015:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2016:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2017:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2018:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2019:mgstr5:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2011:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2012:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2013:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2014:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2015:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth -0.0192284731
## year2016:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2017:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2018:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2019:mgstr8:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2011:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2012:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2013:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2014:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2015:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2016:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2017:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth -0.0860892326
## year2018:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth .
## year2019:mgstr10:bulk_purchase1 Bulk purchase:USA_regionSouth -0.1109367531
## year2011:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
```

```
## year2012:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2013:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2014:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2015:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2016:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2017:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2018:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2019:mgstr4:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2011:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2012:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2013:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2014:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2015:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest -0.0505575490
## year2016:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2017:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2018:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2019:mgstr5:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2011:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2012:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2013:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2014:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2015:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2016:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2017:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2018:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest 0.0003843275
## year2019:mgstr8:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2011:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2012:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2013:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest -0.0472164837
## year2014:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2015:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2016:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2017:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2018:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .
## year2019:mgstr10:bulk_purchase1 Bulk purchase:USA_regionWest .

# final model
model <- lmer(log(ppm) ~ mgstr+bulk_purchase+year+USA_region+(1|USA_region:state), data = streetrx, RE
summary(mod)
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(ppm) ~ mgstr + bulk_purchase + year + USA_region + (1 | USA_region:state)
##    Data: streetrx

##      AIC      BIC   logLik deviance df.resid
##    10607    10737    -5284    10567     4885

## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -4.548 -0.625  0.067  0.571  4.476

## Random effects:
##  Groups           Name        Variance Std.Dev.
##  USA_region:state (Intercept) 0.00112  0.0335
##  Residual                     0.50389  0.7099
## Number of obs: 4905, groups:  USA_region:state, 51

## Fixed effects:
##                              Estimate Std. Error t value
## (Intercept)                   -0.3984     0.2286   -1.74
## mgstr4                        -0.4896     0.0597   -8.19
## mgstr5                        -0.6799     0.0363  -18.71
## mgstr8                        -1.1506     0.1433   -8.03
## mgstr10                       -1.1346     0.0350  -32.43
## bulk_purchase1 Bulk purchase  -0.1681     0.0256   -6.57
## year2011                       0.8669     0.2757    3.14
## year2012                       0.7324     0.2430    3.01
## year2013                       0.7652     0.2350    3.26
## year2014                       0.6555     0.2309    2.84
## year2015                       0.6034     0.2269    2.66
## year2016                       0.6466     0.2256    2.87
## year2017                       0.5845     0.2258    2.59
## year2018                       0.6168     0.2259    2.73
```

```
## year2019                        0.7318     0.2331    3.14
## USA_regionNortheast             0.1846     0.0381    4.85
## USA_regionSouth                 0.0363     0.0311    1.17
## USA_regionWest                  0.1595     0.0349    4.57

### Part 4: Model diagnostics ###
coef(summary(mod))
# random effects: dotplot--estimated random intercepts by state with 95% CI
res <- mod
dotplot(ranef(res, condVar = TRUE))$state
# variance estimates for state-specific variance and residual
VarCorr(res)
## Groups   Name        Std.Dev.
## state    (Intercept) 0.0431
## Residual             0.7108
# residuals vs fitted
plot(res)
# residual distribution
residual <- resid(res)
resid1 <- ggplot() +
  geom_density(aes(x = residual))
resid2 <- ggplot() +
  geom_qq(aes(sample = residual)) +
  geom_qq_line(aes(sample = residual)) +
  coord_equal()

resid1 + resid2

res.inf <- influence(res, "state")
cutoff <- 2/sqrt(length(unique(streetrx$state)))

dfbetas_inf <- round(dfbetas(res.inf), 4)
above_cutoff <- apply(abs(dfbetas_inf) > cutoff, MARGIN = 1, any)
# table of states and DFBETAS for each covariate above cutoff
dfbetas_inf[above_cutoff, ]
## (Intercept) mgstr4 mgstr5 mgstr8 mgstr10
## New York -0.0464 0.2930 0.3984 -0.4136 0.1941
## Nevada 0.3874 -0.1284 -0.0735 -0.1295 -0.1406
## Ohio -0.2834 0.1783 0.0323 0.1129 0.1851
## California 0.0532 -0.3392 -0.4760 0.5029 -0.1635
## Tennessee 0.0487 0.1229 -0.0486 -0.0193 -0.1383
## Colorado -0.0141 -0.3487 0.0375 0.0014 0.1040
## Arizona -0.2349 -0.2251 -0.2226 -0.0336 -0.2435
## Louisiana -0.0217 -0.0753 0.0426 -0.0034 0.0094
## Michigan 0.0584 0.0587 -0.1368 0.1521 -0.0176
## Illinois 0.0293 0.0333 0.1643 0.0651 0.1317
## North Carolina -0.0276 0.1438 0.2841 0.0362 0.2240
## Pennsylvania 0.0224 -0.0054 -0.0879 -0.0594 -0.2144
## New Jersey 0.0061 -0.0488 -0.0628 0.0080 -0.0894
## Oregon -0.0030 -0.0773 -0.0169 0.4626 -0.1396
## Virginia -0.0098 0.5595 0.0966 0.0506 0.1313
## Minnesota -0.0326 0.0741 0.0107 -0.0053 -0.0956
## South Carolina -0.0534 0.1323 0.4141 0.5629 0.3364
## Texas -0.0087 -0.1836 -0.0604 -0.3824 0.1384
## Washington 0.0321 0.1887 0.0008 -0.2738 -0.1294
## Massachusetts -0.0222 0.0752 0.0803 0.0066 0.1361
## Indiana 0.0485 -0.2125 -0.2935 -0.0541 -0.2444
## Connecticut -0.0163 0.0603 0.1191 0.0304 0.0789
## Wyoming 0.0052 0.0202 -0.0096 -0.2912 -0.0270
## bulk_purchase1 Bulk purchase year2011 year2012 year2013 year2014
## New York 0.2829 -0.2293 0.0658 -0.1453 -0.0304
## Nevada 0.0362 -0.3169 -0.2531 -0.3450 -0.3933
## Ohio -0.1712 0.3764 0.2990 0.2666 0.2241
## California -0.2464 0.2368 -0.0018 0.1054 0.0769
## Tennessee 0.0250 -0.0348 -0.0761 -0.0544 -0.0449
## Colorado -0.0978 0.0063 0.0282 0.0330 -0.0184
## Arizona 0.2542 0.2971 0.2081 0.3019 0.2570
## Louisiana -0.0495 0.1731 0.0076 0.0368 0.0165
## Michigan 0.1331 -0.0879 0.1121 0.0531 0.0087
## Illinois -0.0898 -0.0108 -0.1343 -0.1033 0.1546
## North Carolina 0.3314 -0.0002 0.0715 0.0488 0.0197
## Pennsylvania 0.1840 -0.2536 -0.0107 -0.0582 0.0302
```

```
## New Jersey -0.1670 0.0120 -0.0115 -0.0242 0.0009
## Oregon -0.2040 0.0093 0.0597 -0.0192 0.0397
## Virginia -0.1239 0.0012 -0.0193 0.0366 -0.0381
## Minnesota 0.0130 0.0138 0.0255 -0.0290 -0.0009
## South Carolina 0.4299 0.0080 -0.0891 0.0877 -0.0087
## Texas 0.1074 -0.0502 0.0567 -0.0184 -0.1164
## Washington 0.0251 -0.0142 -0.0626 -0.0299 -0.0130
## Massachusetts -0.0053 -0.0012 0.0030 0.0378 -0.0020
## Indiana -0.0850 -0.0124 -0.0400 -0.0102 -0.0024
## Connecticut 0.1399 0.0104 0.0071 0.0229 -0.0184
## Wyoming 0.0403 -0.0025 0.0078 -0.0009 -0.0024
## year2015 year2016 year2017 year2018 year2019 USA_regionNortheast
## New York -0.0391 0.0121 -0.0090 0.0153 -0.0584 -0.6944
## Nevada -0.3736 -0.3820 -0.3772 -0.3746 -0.3502 0.0057
## Ohio 0.2135 0.2937 0.2401 0.2642 0.2604 0.0910
## California 0.0127 -0.0312 -0.0137 -0.0074 0.0182 -0.0147
## Tennessee -0.0355 -0.0443 -0.0363 -0.0464 0.0143 0.0078
## Colorado 0.0422 0.0043 0.0142 0.0139 -0.0326 -0.0139
## Arizona 0.2986 0.2583 0.2625 0.2556 0.2778 -0.0117
## Louisiana -0.0019 0.0105 0.0144 0.0195 0.0759 0.0200
## Michigan 0.0270 -0.0101 -0.0022 0.0127 -0.0779 -0.3575
## Illinois -0.0041 0.0059 0.0110 0.0132 -0.0308 -0.3531
## North Carolina 0.0019 -0.0422 -0.0041 -0.0001 -0.0227 -0.0161
## Pennsylvania 0.0081 -0.0052 0.0071 -0.0421 0.0188 0.9488
## New Jersey -0.0161 0.0055 -0.0001 0.0167 0.0172 -0.4631
## Oregon 0.0268 0.0061 0.0368 0.0056 0.0098 -0.0023
## Virginia -0.0246 0.0103 -0.0020 -0.0404 0.0149 0.0116
## Minnesota -0.0186 -0.0168 -0.0115 0.0093 0.0013 0.2928
## South Carolina -0.0083 -0.0047 0.0066 -0.0193 -0.0328 -0.0154
## Texas 0.0078 0.0091 0.0288 0.0008 -0.0585 0.0051
## Washington -0.0118 -0.0139 -0.0426 -0.0298 -0.0252 0.0017
## Massachusetts 0.0166 0.0011 -0.0034 0.0092 0.0107 0.5294
## Indiana 0.0073 0.0017 0.0037 -0.0032 -0.0268 -0.0482
## Connecticut 0.0066 -0.0051 -0.0049 -0.0011 0.0051 -0.3631
## Wyoming -0.0022 0.0006 -0.0054 -0.0056 -0.0022 0.0011
## USA_regionSouth USA_regionWest
## New York 0.0340 0.0282
## Nevada -0.0035 0.0628
## Ohio 0.0861 0.0813
## California -0.0167 0.2065
## Tennessee 0.2807 0.0063
## Colorado -0.0219 -0.0636
## Arizona 0.0009 -0.1965
## Louisiana 0.2990 0.0380
## Michigan -0.4260 -0.4075
## Illinois -0.4216 -0.3781
## North Carolina -0.1285 -0.0138
## Pennsylvania 0.0405 -0.0154
## New Jersey 0.0343 0.0294
## Oregon 0.0161 0.3231
## Virginia -0.1801 -0.0187
## Minnesota 0.3665 0.3279
## South Carolina 0.0329 -0.0191
## Texas -0.0993 0.0031
## Washington 0.0175 -0.2998
## Massachusetts 0.0162 0.0025
## Indiana -0.0677 -0.0551
## Connecticut 0.0170 0.0070
## Wyoming 0.0002 -0.0253

# Extract DFBETAS from the influence object
dfbetas_values <- as.data.frame(dfbetas(res.inf))
# Add a column for the state
dfbetas_values$State <- rownames(dfbetas_values)
# Convert from wide to long format for plotting
dfbetas_long <- tidyr::gather(dfbetas_values, "Parameter", "DFBETAS", -State)
# Create the plot for DFBETAS
ggplot(dfbetas_long, aes(x = DFBETAS, y = State, color = Parameter)) +
  geom_point(size = 1) +
  theme_minimal() +
  labs(title = "DFBETAS␣for␣Each␣State",
       x = "DFBETAS",
```

```
       y = "State") +
  geom_vline(xintercept = c(-cutoff, cutoff), linetype = "dashed")

# Compute Cook's distance
cooks_distance <- cooks.distance(res.inf)

# Create a data frame with state and corresponding Cook's distance
cooks_df <- data.frame(State = unique(streetrx$state),
                       CooksDistance = cooks_distance)

# Calculate mean Cook's distance for each state and arrange in descending order
cooks_df <- cooks_df %>%
  group_by(State) %>%
  summarize(MeanCooksDistance = mean(CooksDistance, na.rm = TRUE)) %>%
  arrange(MeanCooksDistance)

# Convert State to a factor with levels in the order of MeanCooksDistance
cooks_df$State <- factor(cooks_df$State, levels = cooks_df$State)

# Plot Cook's distance
ggplot(cooks_df, aes(x = MeanCooksDistance, y = State)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Mean Cook's Distance for Each State",
       x = "Mean Cook's Distance",
       y = "State") +
  geom_vline(xintercept = 4/length(cooks_distance), linetype = "dashed", color = "red")
```

6.2 Additional plots

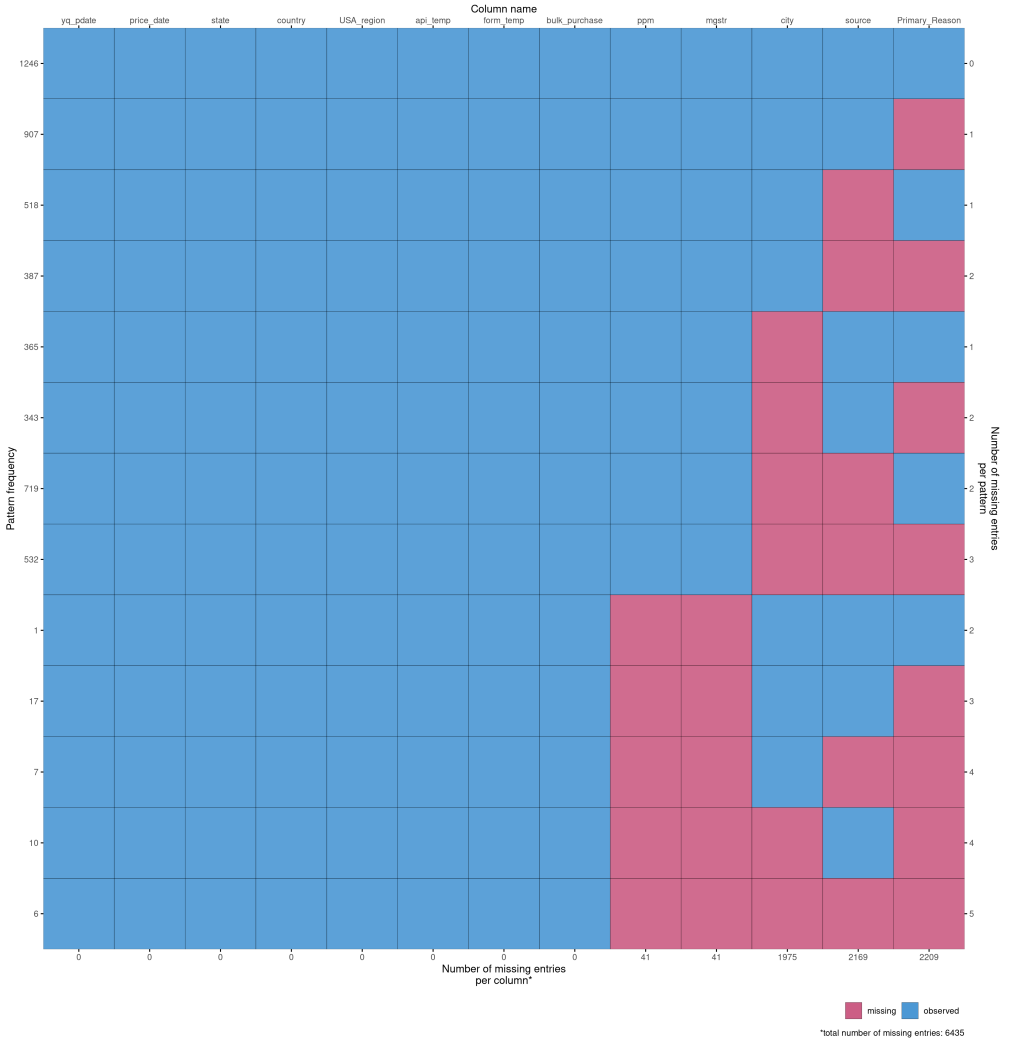

Figure 7: Missing patterns for each variable

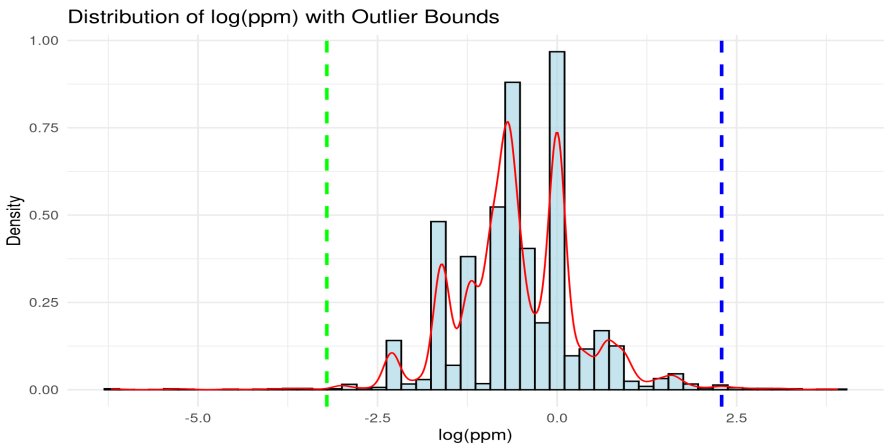

Figure 8: Threshold for removing outliers in log(`ppm`)

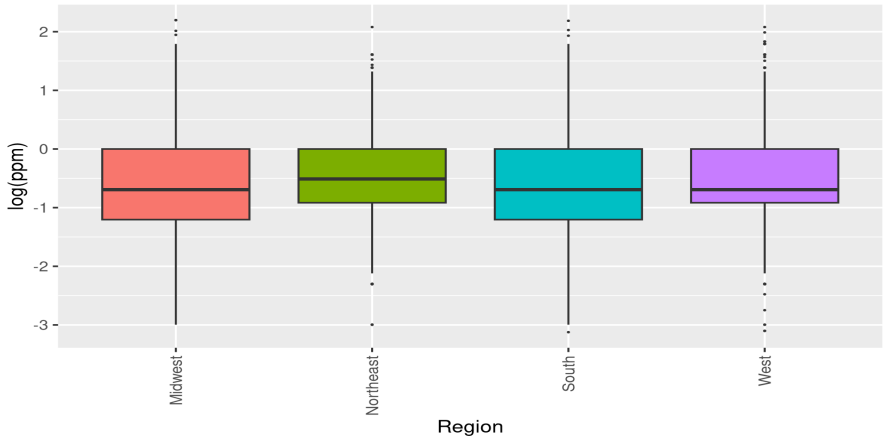

Figure 9: log(`ppm`) by region

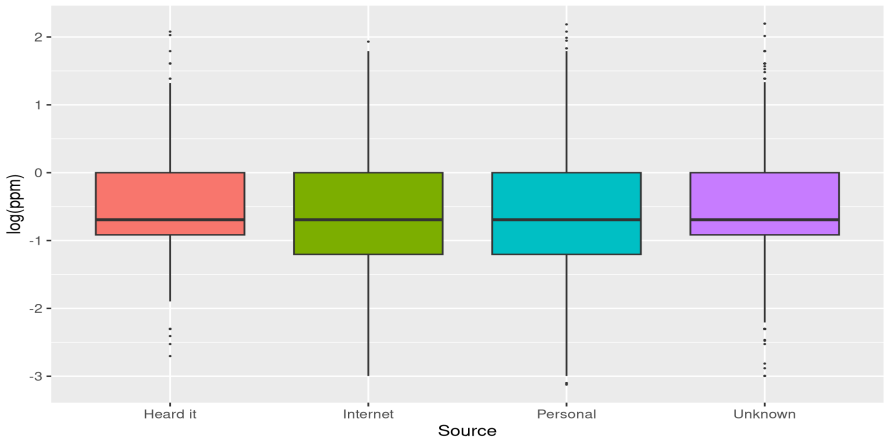

Figure 10: log(`ppm`) by `source`

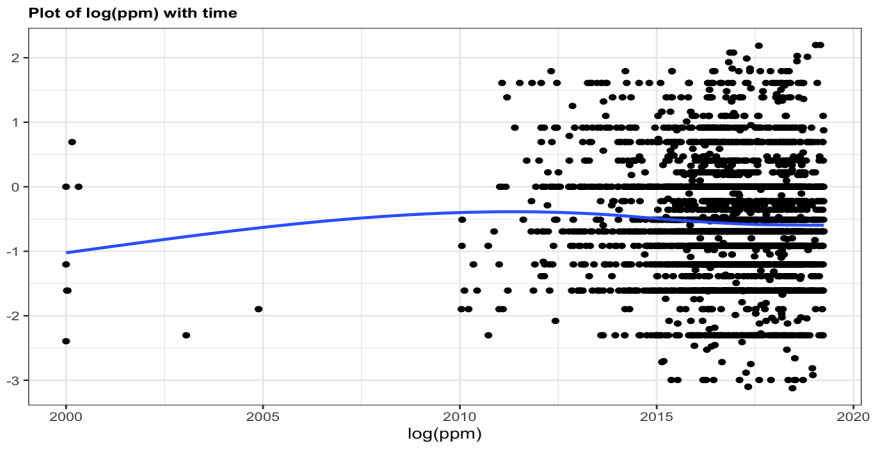

Figure 11: log(`ppm`) by time

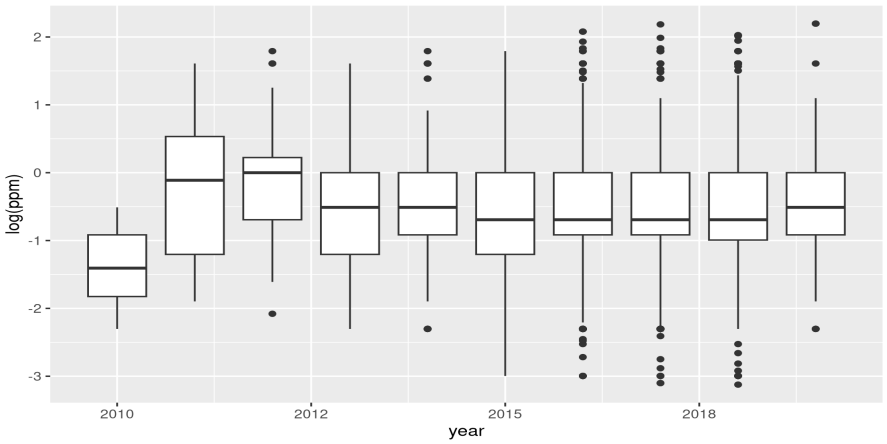

Figure 12: log(ppm) by year

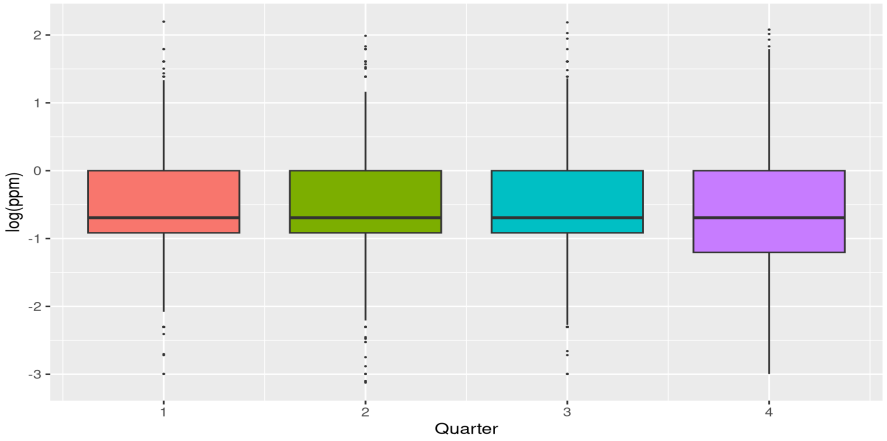

Figure 13: log(ppm) by quarter

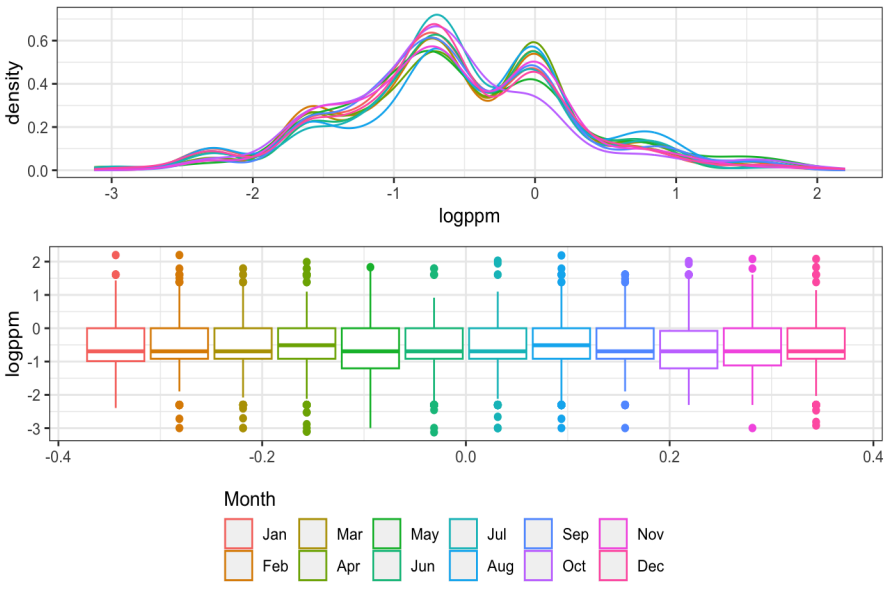

Figure 14: log(ppm) by month

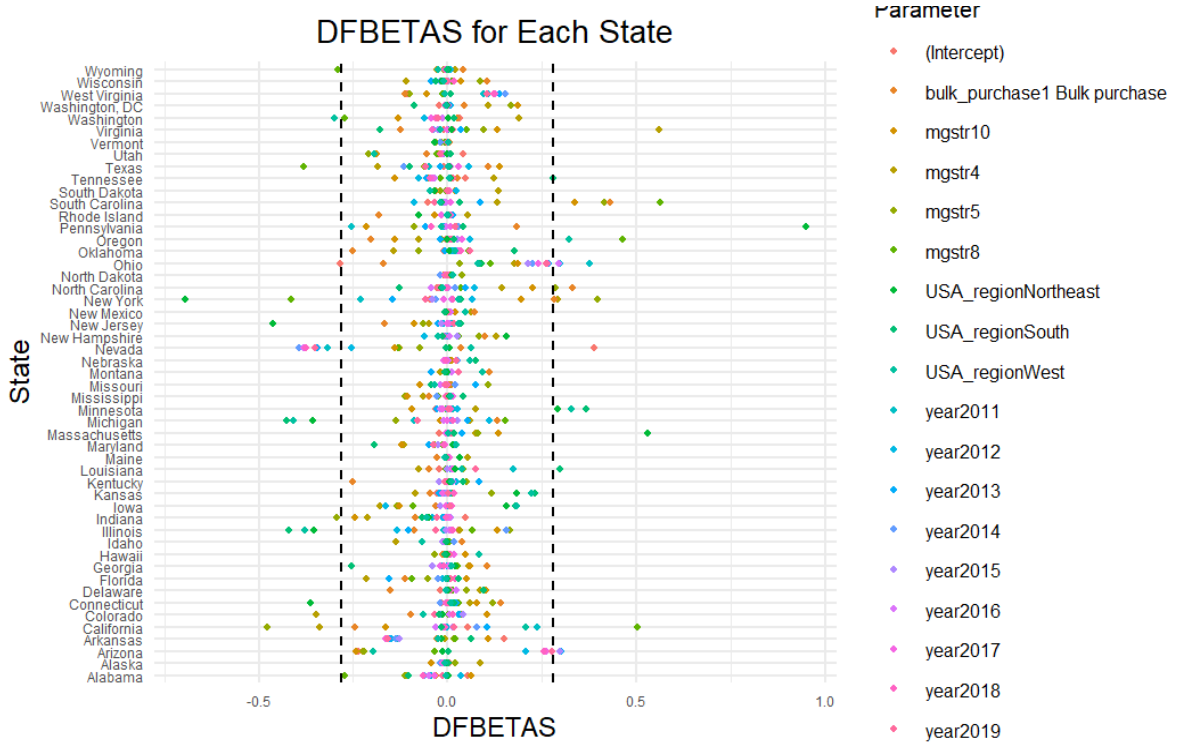

Figure 15: DFBETAS of each parameter for each state (vertical dash lines are the chosen cutoffs)

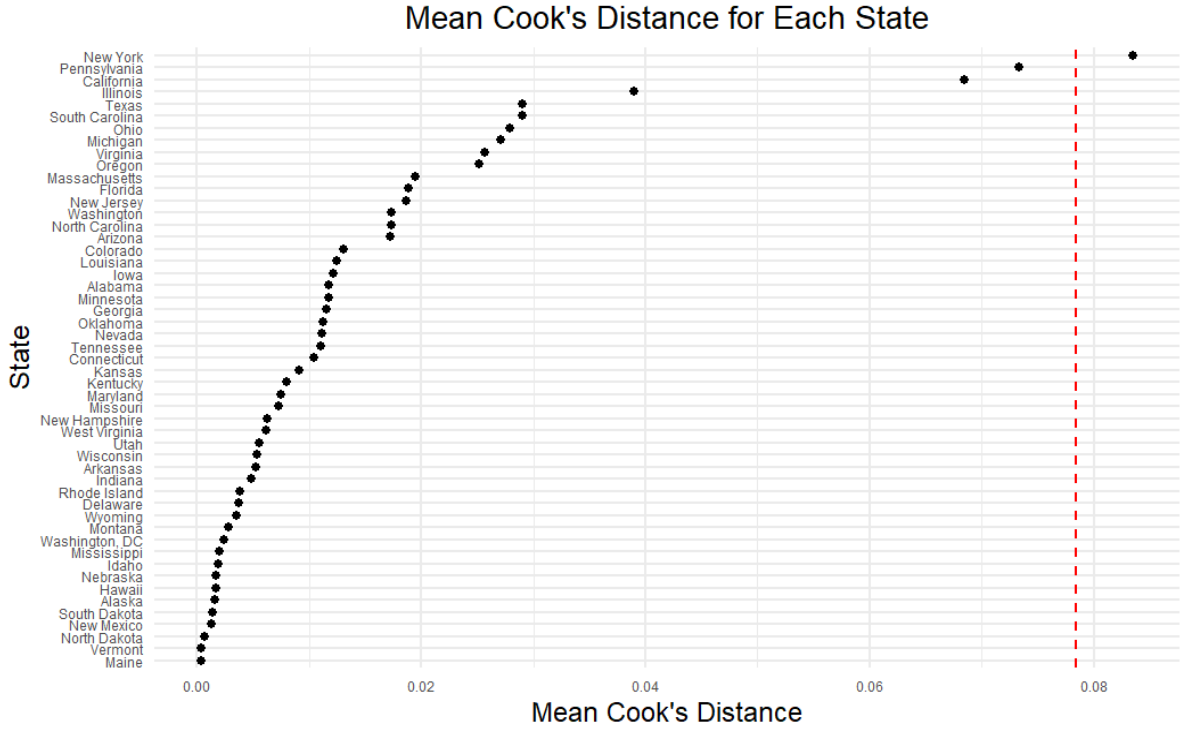

Figure 16: Cook's distance by state (vertical dash line is the chosen cutoff)