STA610 Case Study 1

Emily Gentles (Presenter) Weiyi Liu (Writer) Jack McCarthy (Programmer)
Qinzhe Wang (Coordinator & Checker)

17 October, 2021

Introduction

Prescription opioid abuse has recently become an epidemic in the United States. The price of illicit prescription opioids indicates the supply-demand relationship of the drug. This case study aims to explore the relationship between the unit price of drugs and other factors such as the quantity purchased, the location of the transaction, and strength of the drug. More specifically, our group's interest is to explore the factors related to the cost per milligram and the heterogeneity in the region. The dataset we will be using is provided by StreetRx, a reporting tool for people at large to anonymously report the price they paid or heard for diverted prescription drugs. Our drug of interest is Morphine which is used to "relieve moderate to severe pain and may be habit-forming," especially with prolonged use (MedlinePlus).

Data Cleaning & EDA

Missing Values

The subset of the StreetRx dataset pertaining to Morphine contains 9,268 observations with 13 variables. There are 13,443 empty cells, including both missing values and blank entries. To maintain the statistical power and avoid bias, our group decided to recode both the empty cells and "0 Reporter did not answer this question" in `Primary_Reason` (5061 in total) as "8 Prefer not to answer" and recode the empty cells in `source` (3942 in total) as "Blank" because of the high missing rates. Then, we removed other rows with missing values.

Additionally, we removed non-positive price values as well as price values greater than 10. Since the data is self-reported, these extremely expensive prices are likely due to users misunderstanding the system and reporting total price instead of unit price. The number of observations is now 5,582.

Response Variable: Price per milligram

Figure 1: Distribution of morphine ppm

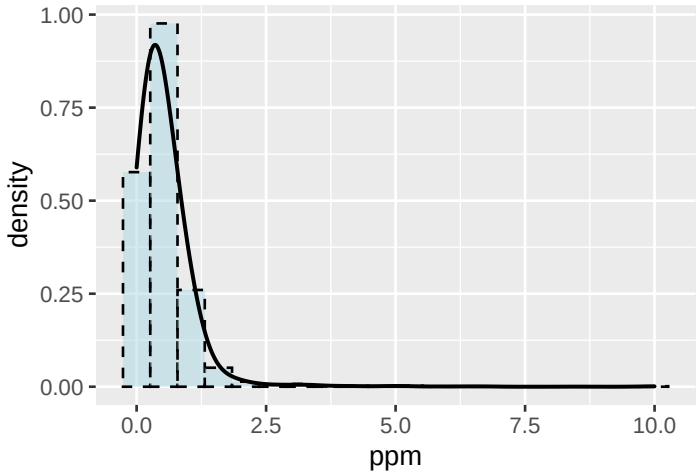
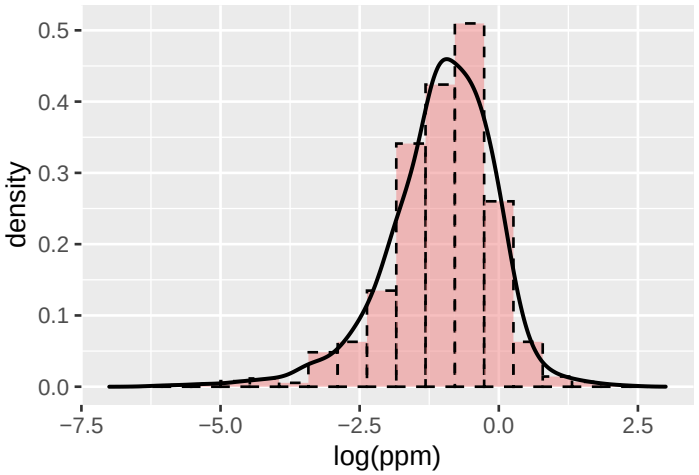


Figure 2: Distribution of morphine log(ppm)



Whether we fit a hierarchical model or linear regression, the response variable should be normally distributed. Although the normality assumption pertains to the conditional distribution of our response variable, it's still beneficial to check the assumption for the marginal distribution as a very skewed marginal distribution could persist and affect the model's resulting conditional distribution. From the histogram on the left, the distribution of `ppm` is clearly right-skewed. Since `ppm` is strictly non-negative, a log transformation may be appropriate. We can see that the distribution of `log(ppm)`, given above, appears to be much closer to the desired normal distribution.

Grouping Variable: city, state, and region

Since we want to analyze the heterogeneity in pricing by location, we have three choices of grouping variables, `city`, `state`, and `USA_region`.

City

There are 1642 unique `city` values, and many cities have small sample size (i.e. less than 5 observations). We decide not to use `city` as the grouping variable (see appendix).

State

As for the state, we examined the sample sizes in each group and decided to filter out Puerto Rico and Vermont because they have less than 5 observations.

Figure 3: Group mean vs. sample size

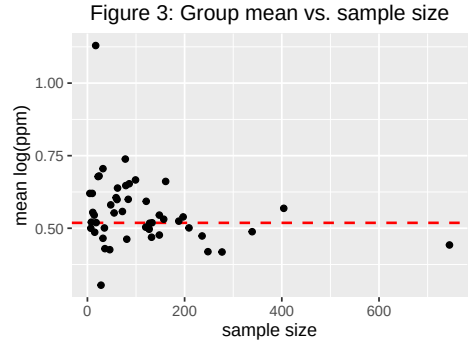
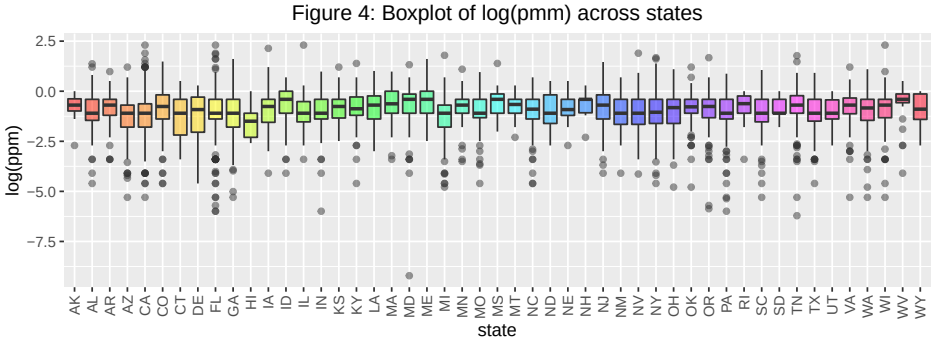


Figure 4: Boxplot of log(pmm) across states



We then inspect the state-level differences more closely by plotting the group-level means against the sample sizes. We observed that the within-state means for states with smaller sample sizes vary a lot, while the within-state means for states with higher sample sizes in general adhere more closely to the grand mean. This is conducive to the borrowing of information between states with a hierarchical model. From the above boxplot of `log(ppm)` against `state`, it is also evident that the `log(ppm)` distributions differ across states. This indicates

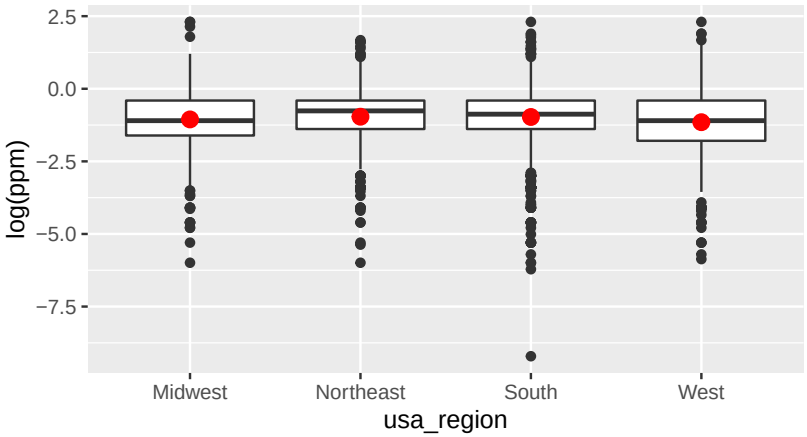
the potential state-level differences in drug prices. Therefore, we decide to use state as our grouping variable at this stage.

Region

From the boxplot we see that the `log(ppm)` distributions differ slightly across regions, though not as much as across states. We may also consider using region as the grouping variable.

	usa_region	n	mean
1	Midwest	1168	−1.056
2	Northeast	674	−0.962
3	South	1953	−0.972
4	West	1773	−1.151

Figure 5: Boxplot of log(pmm) across regions



Date (price_date)

As for the `price_date`, we noticed some observations are prior to the establishment of StreetRx, which are likely incorrect inputs. We dropped the observations before 2010. For the remaining observations, we came up with two ways of data cleaning. The first choice is to choose a starting date and convert the feature as the date differences (`date_diff`) from that starting date. The second choice is to split this date variable into two components, `year` and `quarter`, to explore the trend of unit drug price over time and the seasonality.

Our visualizations suggested there is no clear indication that the log value of per milligram price of morphine varies along with `date_diff`. However, for different `year` and `quarter`, the `log(ppm)` value varies slightly (see appendix).

Bulk_purchase & Source

Figure 8: log(pmm) vs. bulk purchase

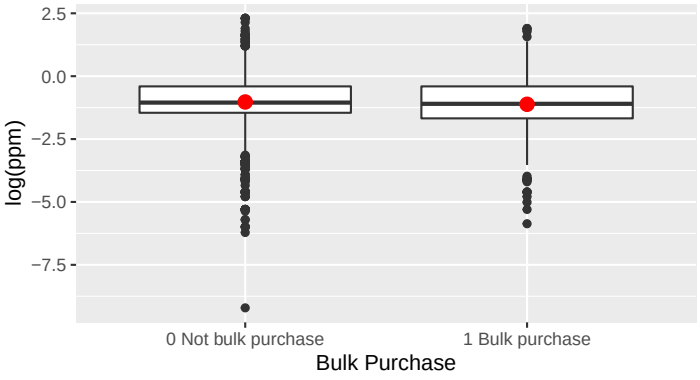
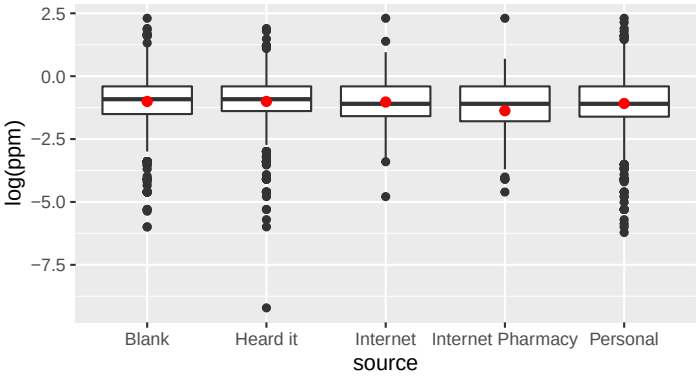


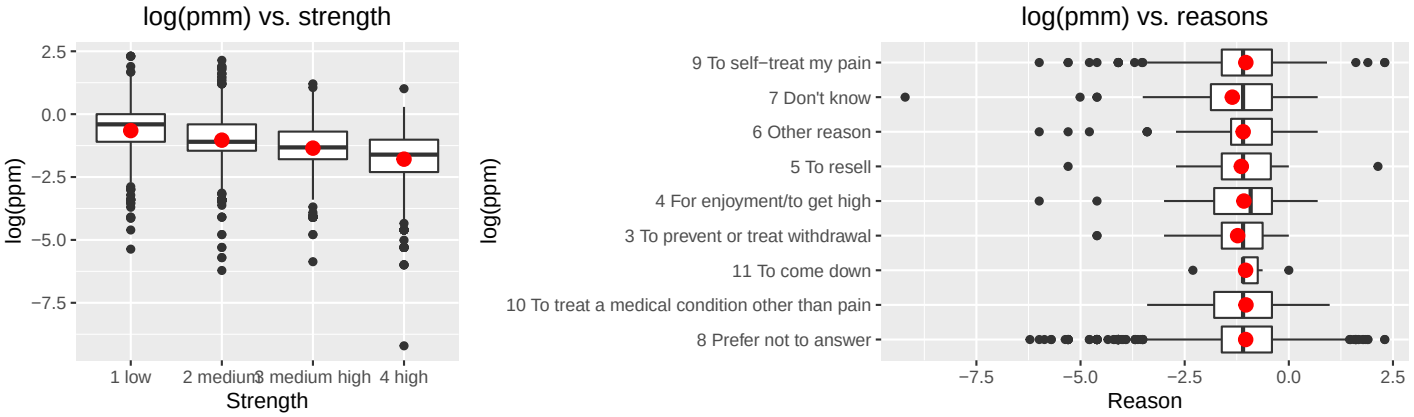
Figure 9: log(pmm) vs. source



There is no need to conduct any data cleaning on `bulk_purchase`. From the boxplot we see that there is a slight trend that the drug price may be lower if purchased in bulk. Therefore, `bulk_purchase` might be a potential predictor.

For the feature `source`, we have recoded the missing value as "Blank" and the name of websites as "Internet". We also dropped the only observation whose `source` is "Drug Forum". The boxplot shows that the `log(ppm)` value varies among different sources (see appendix).

Dosage Strength & Primary Reason



From the scatter plot of `log(ppm)` against `mgstr`, there is a slight trend that the larger the dosage strength, the smaller the per milligram price. We have also noticed that `mgstr` only takes 16 discrete values. Therefore, we decided to transform it into 4 levels ("low", "medium", "medium high", and "high") based on the 0.25, 0.5, and 0.75 quantiles of `mgstr`. From the boxplot, the trend that the `log(ppm)` values decrease as the dosage strength increases is more clear when using these new levels.

For `primary_reason`, we have converted the empty cells and "0 Reporter did not answer this question" to "8 Prefer not to answer". The `log(ppm)` value varies among different reasons for purchasing morphine (see appendix).

Interaction

We also inspected how predictors interact with each other, i.e. whether the effect of one predictor on the unit price of morphine is influenced by any other predictor. We found that there might by potential interaction among `dosage strength`, and `quarter`, or `quarter`, and `primary reason` (see appendix). In the modeling part, we will check whether there are interactions in a more formal way.

Model

Initial Model & Model Selection

The goal of our analysis is to investigate factors related to the per milligram price of morphine and explore heterogeneity in pricing by location. As discussed in the EDA part, we do not have enough data to estimate the effects at the city level. Meanwhile, the drug prices do not seem to change significantly across regions. Thus, the state variable is the prefered choice of accounting for location. Since many states have relatively small sample sizes, a hierarchical model allows us to borrow information across states.

Comparing three full models with different grouping variables, the AIC and BIC score also suggest choosing `state` as the group-level variable.

Our baseline model incorporates only the state-level random intercepts. For other individual-level predictors, we add one variable to the model each time and use both the Likelihood Ratio test and the BIC score to determine whether it should be added. The LRT is designed for nested models while the BIC score considers both the likelihood and the model complexity and gives a more general sense of model performance while also

Table 1: AIC and BIC for different grouping variables

Grouping	AIC	BIC
City	15408.58	15428.46
State	15354.88	15374.76
Region	15400.48	15420.36

being consistent. Tables 1 and 2 display the results of model selection. We also used the full model as a starting point to perform stepwise backward elimination with the results agreeing with the previous model selection method. (See appendix).

Our final model incorporates the grouping variable `state` and the individual level predictors `mgstr` (recoded as 4 levels), as well as `bulk_purchase`, `quarter`, and `source`.

Table 2: Forward model selection

Model	LRT.p.value	BIC
(1|state)		15374.76
(1|state) + mgstr2	0	14615.63
(1|state) + mgstr2 + bulk_purchase	2e-04	14610.01
(1|state) + mgstr2 + bulk_purchase + year	0.1079	14673.21
(1|state) + mgstr2 + bulk_purchase + quarter	0.0213	14626.19
(1|state) + mgstr2 + bulk_purchase + date_diff	0.1844	14616.87
(1|state) + mgstr2 + bulk_purchase + quarter + source	7e-04	14641.45
(1|state) + mgstr2 + bulk_purchase + quarter + source + primary_reason	1	14681.42

Interactions

From the EDA, we see some potential interactions between predictors (see appendix). Here, we also tried to incorporate all possible two-way interactions into the model (one at a time), but non of them seem to pass the LRT test or improve the model BIC score (see appendix). To control the model complexity, we did not try any three-way or more complex interaction terms. Therefore, we decided not to add any interaction term into the final model.

Final Model

Our final model is

$$log(y_{ij}) = \beta_0 + b_{0j} + \beta_1 M_{ij} + \beta_2 B_{ij} + \beta_3 Q_{ij} + \beta_4 S_{ij} + \epsilon_{ij}$$

$$b_{0j} \overset{iid}{\sim} \mathcal{N}(0, \tau^2) \perp \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

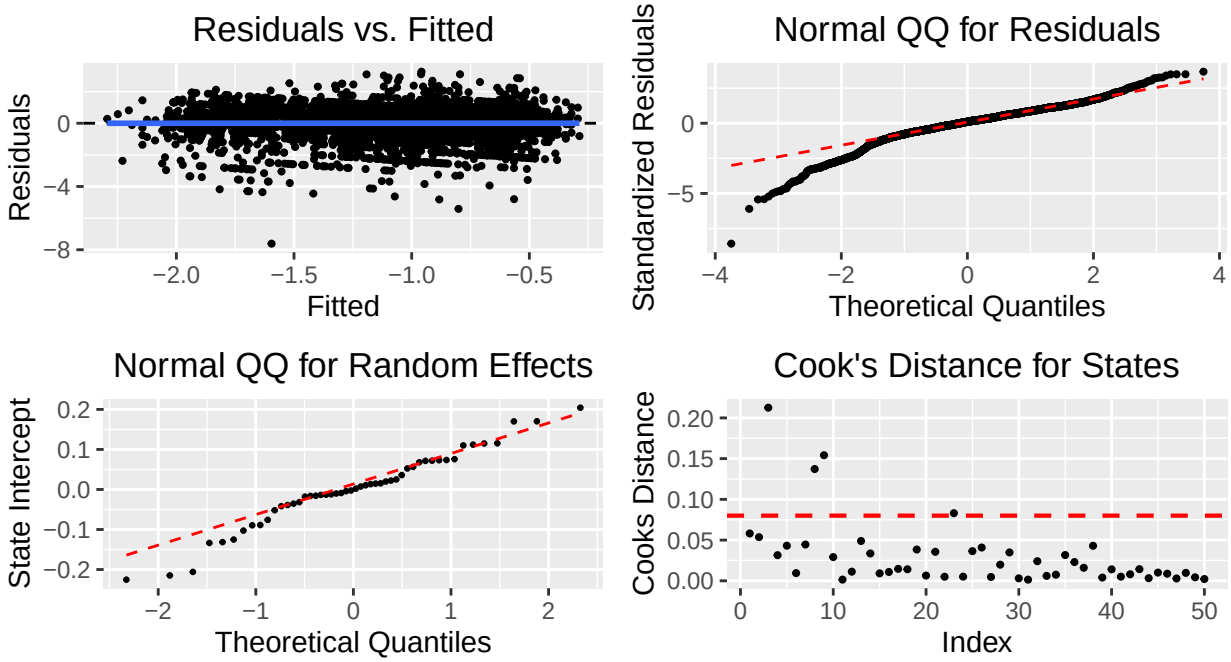
The response variable and predictors are defined as:

- $y_i j$: Per milligram price of morphine for individual i in state j
- M_{ij}: Dosage strength in mg of the units purchased, factored into 4 levels
- B_{ij}: Bulk purchase, an indicator for whether 10+ units were purchased at once
- Q_{ij}: Quarter of the reported purchase
- S_{ij}: Source of information (including first-hand and second-hand sources)

Model Assumptions

- There is a linear relationship between the dependent variable `log(ppm)` and the predictors
- y_{ij}'s are independent
- The variance of y_{ij} in each group should be the same
- $\epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Model Diagnostics



- `Residual vs. Fitted plot`: The residuals are spread equally around the horizontal line, indicating there is no non-linear relationship. Besides, the equal variance assumption is met.
- `Normal QQ plot for residuals`: The normality assumption is slightly met since our residuals adhere around the diagonal line representing normality but have heavy tails on both sides. We also have one data point that deviates severely from the diagonal line.
- `Normal QQ plot for Random Effects`: We can accept the random effects are normally distributed. But we still have three outliers.
- `Cook's Distance`: We have 3 highly influential states (Florida, Pennsylvania, and California) whose Cook's distance exceeds the $\frac{4}{n}$ cutoff, where n denotes the number of states.

To address the violated assumptions, we tried to remove the data point with the lowest residual and the influential groups. By removing the influential observation, the normality of the residuals improved a little bit, and the dots aligned more tightly to the diagonal line. However, removing the influential states does not drastically improve the normality of the residuals (see appendix). Moreover, the influential states have a considerable sample size (1382 observations). Therefore, we decide only to drop the individual level outlier but keep all the groups.

Table 3: Estimates of fixed effects

	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.6346	0.5301	0.0393	296.1819	-16.1290	0.0000
quarter2	0.0854	1.0891	0.0321	5550.5948	2.6578	0.0079
quarter3	0.0841	1.0877	0.0332	5555.0726	2.5349	0.0113
quarter4	0.0844	1.0881	0.0341	5551.1650	2.4759	0.0133
sourceHeard it	0.0633	1.0653	0.0335	5556.1197	1.8904	0.0588
sourceInternet	-0.0041	0.9959	0.0625	5555.5822	-0.0656	0.9477
sourceInternet Pharmacy	-0.3227	0.7242	0.1016	5548.9347	-3.1743	0.0015
sourcePersonal	-0.0398	0.9609	0.0281	5557.7473	-1.4157	0.1569
mgstr22 medium	-0.3816	0.6827	0.0279	5549.1951	-13.6631	0.0000
mgstr23 medium high	-0.7000	0.4966	0.0365	5554.7006	-19.1836	0.0000
mgstr24 high	-1.1197	0.3264	0.0420	5559.7107	-26.6889	0.0000
bulk_purchase1 Bulk purchase	-0.1141	0.8922	0.0296	5557.9880	-3.8488	0.0001

Conclusion

Fixed Effects

- Quarter (baseline: Quarter1): Compared with quarter 1, holding all other predictors unchanged, purchasing the morphine in quarter 2, the per milligram price of the drug will increase by a multiplicative effect of $e^{0.0853} = 1.0891$ (about 8.91%). Similarly, if the drug is purchased in quarter 3 or 4, the drug price will increase by 8.77% and 8.81%, respectively.
- Source (baseline: Blank): Compared with an unknown source, holding all other predictors unchanged, the per milligram drug price heard from other people will increase by a multiplicative effect of $e^{0.0633} = 1.0653$ (about 6.53%). Similarly, the price information obtained from the internet, internet pharmacy, or personal purchase will decrease by 0.41%, 27.58%, and 3.91%, respectively.
- Dosage Strength (baseline: Low): Compared with low dosage strength, holding all other predictors unchanged, the per milligram price of morphine will decrease by a multiplicative effect of $e^{-0.3816} = 0.6827$ (about 31.73%) if it has medium dosage strength. Similarly, if the dosage strength is medium-high or high, the drug price will decrease by 50.34% and 67.36%, respectively.
- Bulk Purchase (baseline: Not bulk purchase): Compared with non-bulk purchase, holding all other predictors unchanged, the unit price of morphine will decrease by a multiplicative effect of $e^{0.1141} = 0.8922$ (about 10.78%).

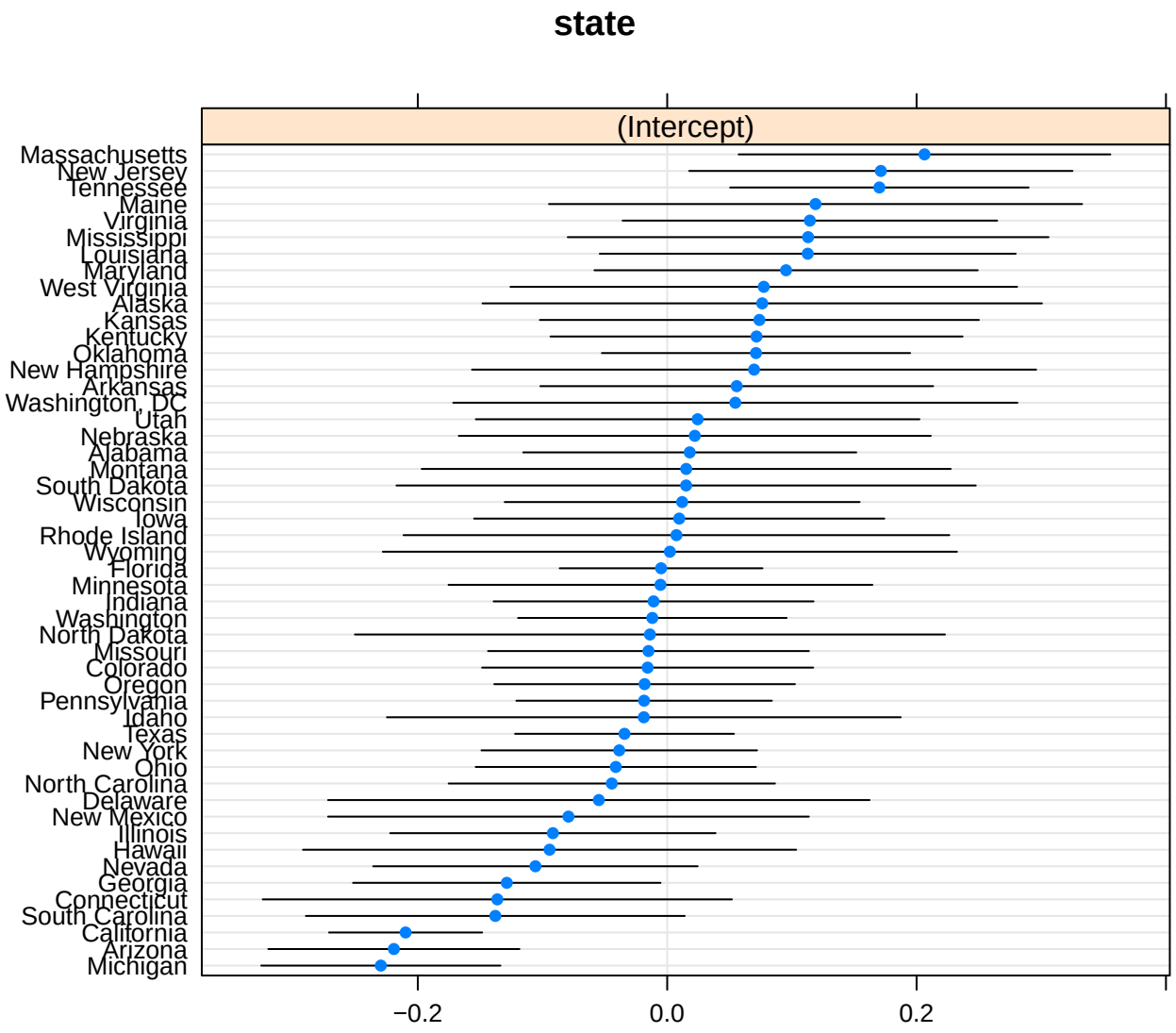
Random Effects

Table 4: Estimates of random effects

	τ^2	σ^2
Estimate	0.0161	0.7772

The estimated across-state variance is $\hat{\tau^2} = 0.0161$, which also describes the variation attributed to the random intercept. The estimated within-state variance is $\hat{\sigma^2} = 0.7772$, which describes the unexplained variation. The estimated interclass correlation is $\frac{\hat{\tau^2}}{\hat{\tau^2}+\hat{\sigma^2}} \approx 0.02$. Therefore, we have little correlation within the same state.

From the random intercepts plot below, we can see that states have different bases per milligram morphine prices. The prices ranges from $e^{-0.2297} = 0.7948$ (Michigan) to $e^{0.2063} = 1.2292$ (Massachusetts). These estimates are based on the baseline condition of all other predictors, which are purchasing in quarter 1, from an unknown source, with low dosage strength, and not purchased in bulk.



Strength and Limitations

Strength: We have collected a considerable amount of data from StreetRX, which allows us to account for potential clustering and exploring heterogeneity in pricing by location (states). Even though some states have small sample sizes, a hierarchical model helps mitigate the issue. We have access to many predictors, which allows a well-rounded analysis of the potential factors that may affect the drug price.

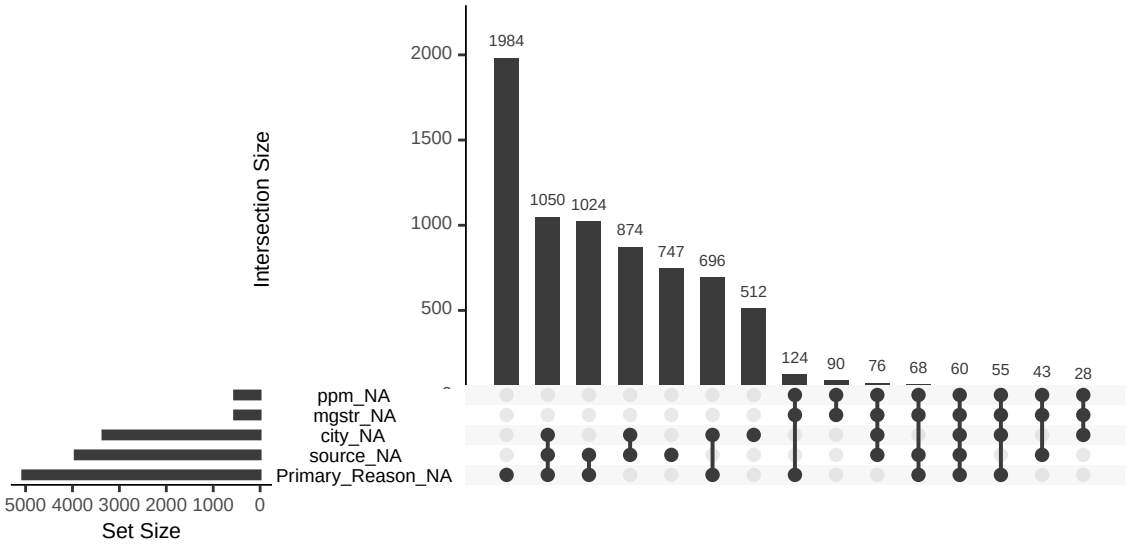
Limitations: Firstly, StreetRX provides only self-reported data, which is likely messy, biased, and lacking credibility. Although we borrow information across states via a hierarchical model, states with small sample sizes are still problematic. The within-state variance σ^2 is much larger than the across-state variance τ^2, indicating that there is still much within-state variation left unexplained. This suggests that the per milligram price of morphine may depend on factors not on the state level. Having access to more predictors may help explain these variations and improve the model performance.

Appendix

Additional tables and figures

Data Cleaning & EDA

Missing Values



Grouping Variable: city, state, and region

State

Date (price_date)

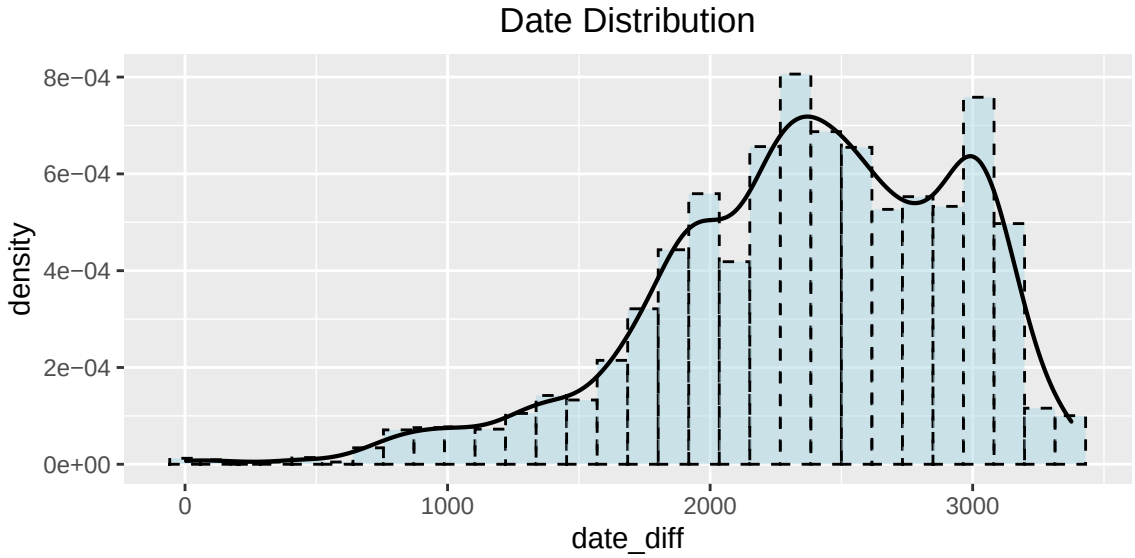
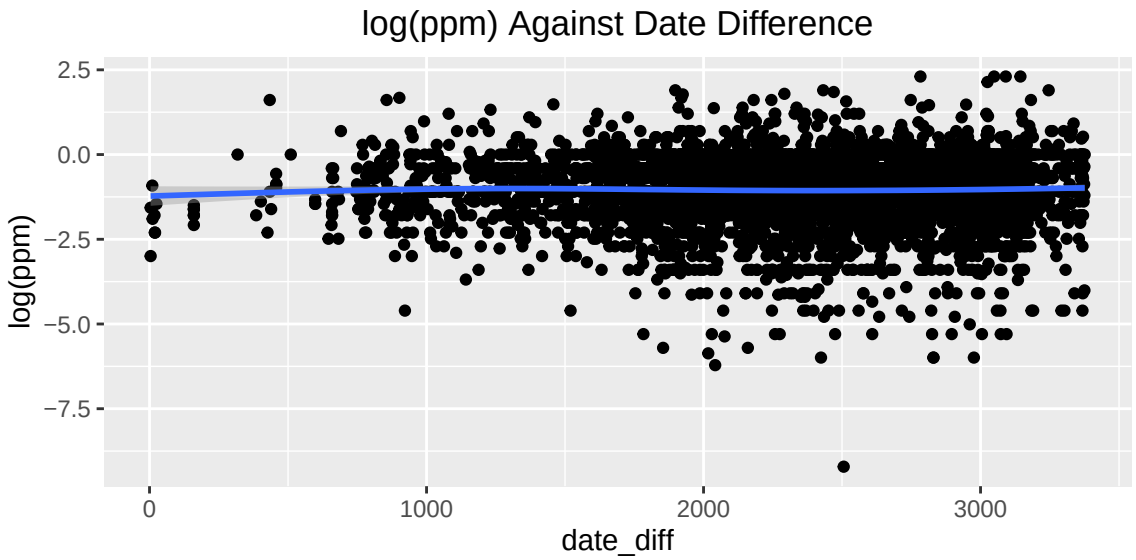
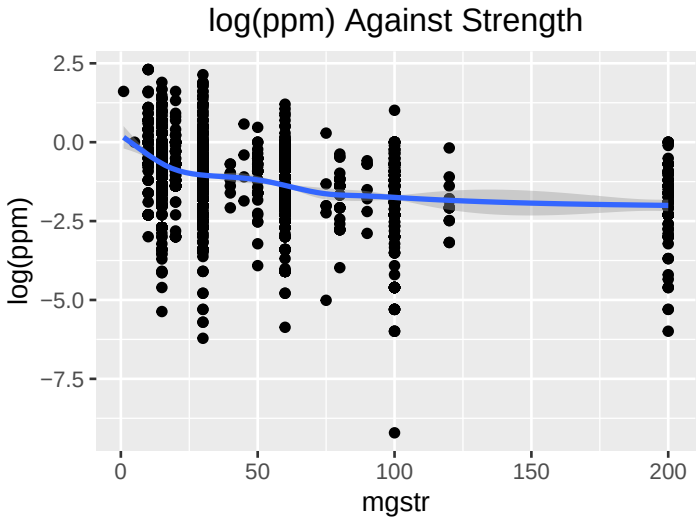


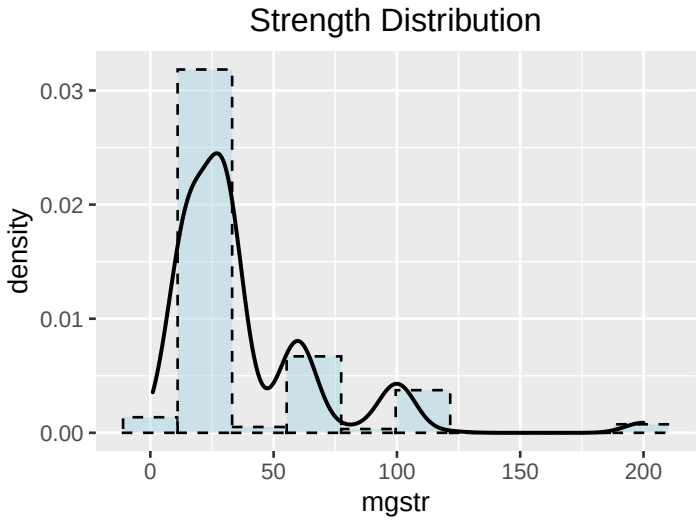
Table 5: Quantile of mgstr

	mgstr
25%	15
50%	30
75%	60

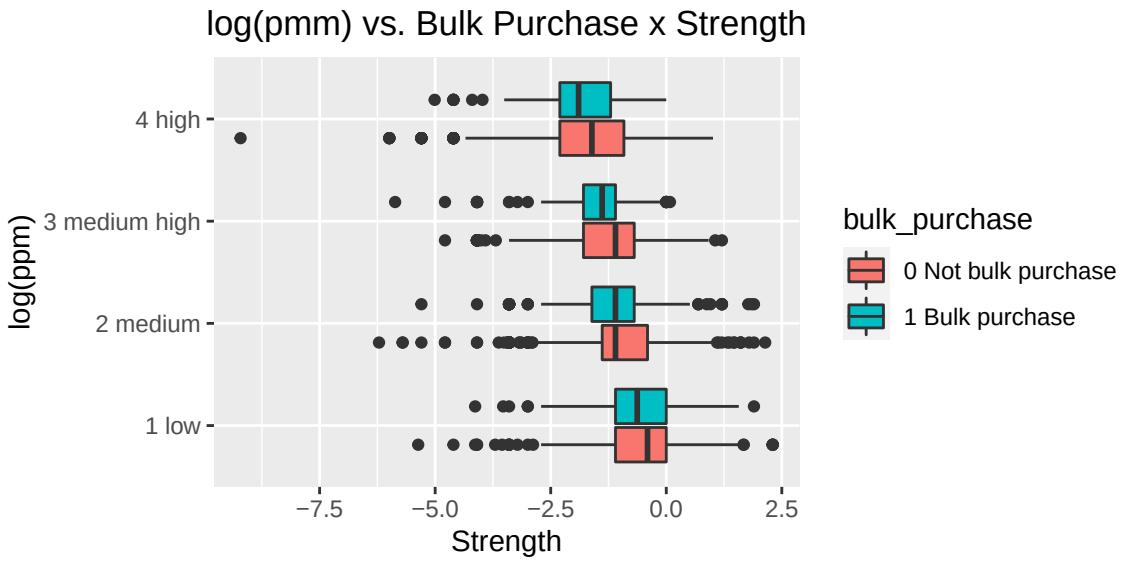


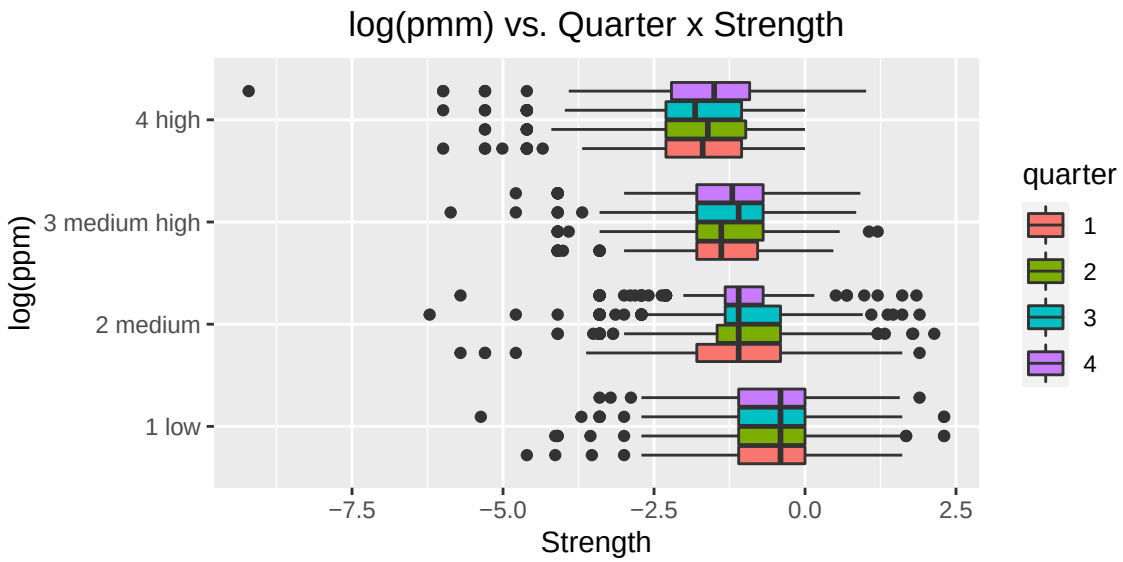
Dosage Strength & Primary Reason

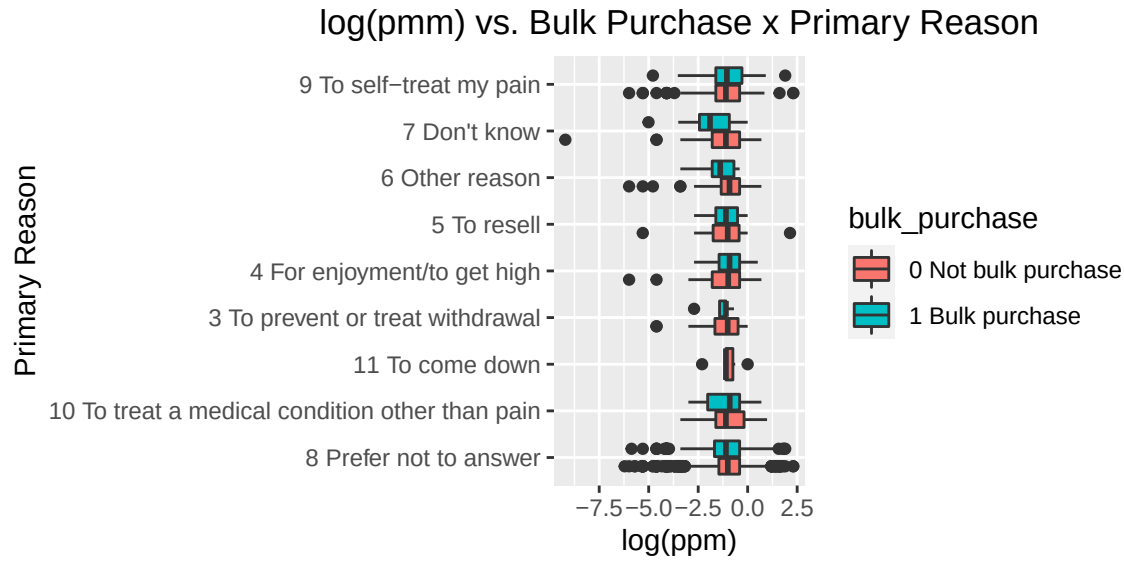


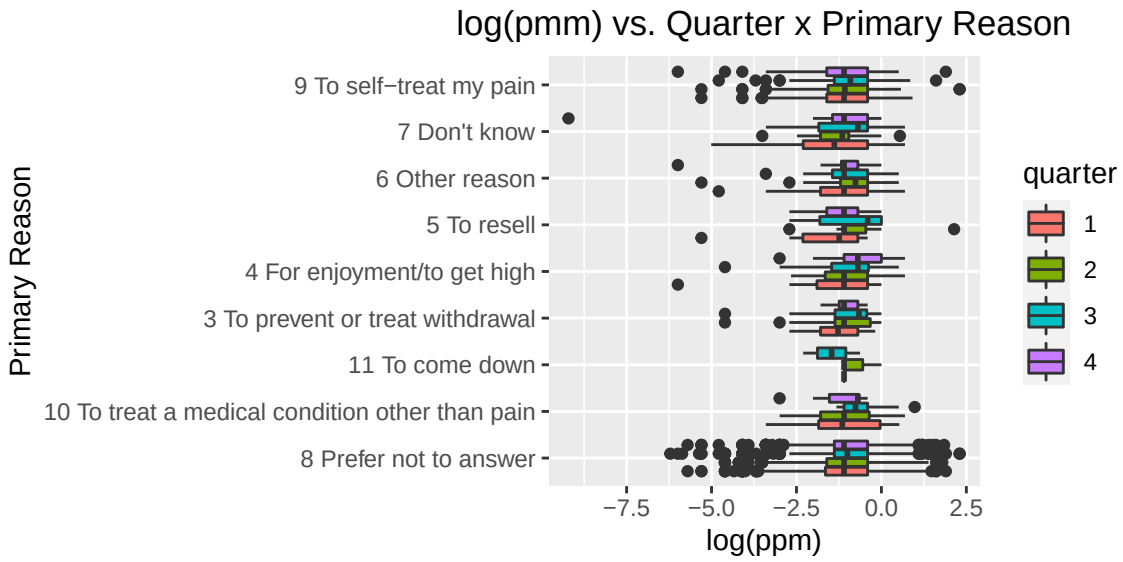


Interaction Plots









Model

Initial Model & Model Selection

Table 6: AIC and BIC for different grouping variables (full models)

Grouping	AIC	BIC
City	14627.21	14931.96
State	14577.10	14881.85
Region	14610.16	14914.91

Interactions

Model Diagnostics

Diagnostic plots for model_g2 (drop one observation with the lowest residual in model_g)

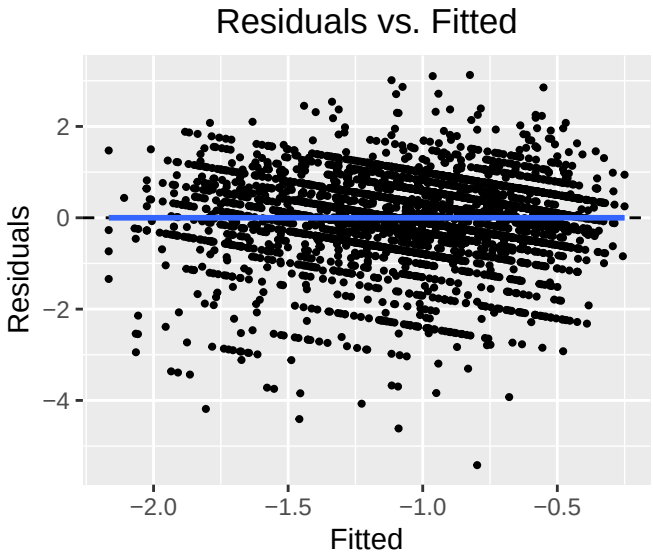
Diagnostic plots for model_g3 (drop influential states)

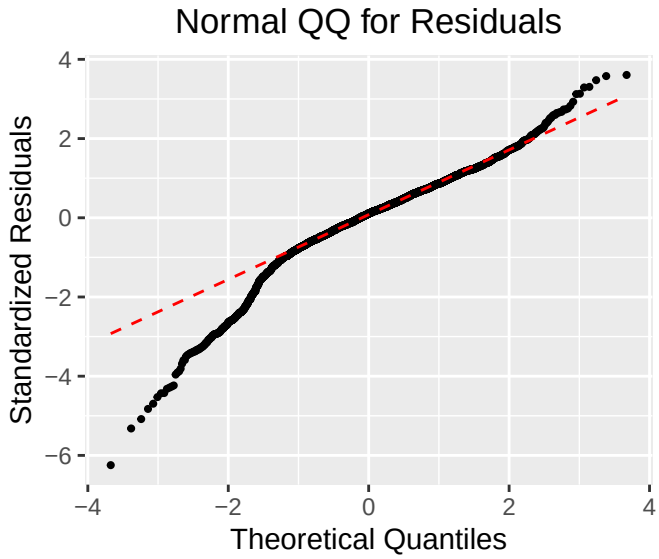
Table 7: Stepwise backward elimination results

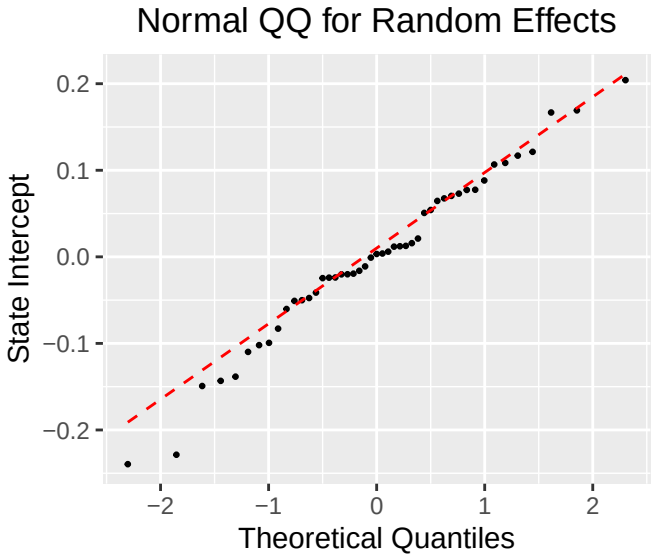
	Eliminated	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
date_diff	1	0.0173	0.0173	1	5560.078	0.0221	0.8817
year	2	8.1386	0.9043	9	5549.587	1.1558	0.3191
primary_reason	3	11.3847	1.4231	8	5552.196	1.8156	0.0693
quarter	0	7.1468	2.3823	3	5554.311	3.0310	0.0282
mgstr2	0	670.1758	223.3919	3	5554.045	284.2265	0.0000
bulk_purchase	0	12.1690	12.1690	1	5560.929	15.4829	0.0001
source	0	24.8908	1.3828	18	5550.635	1.7594	0.0243

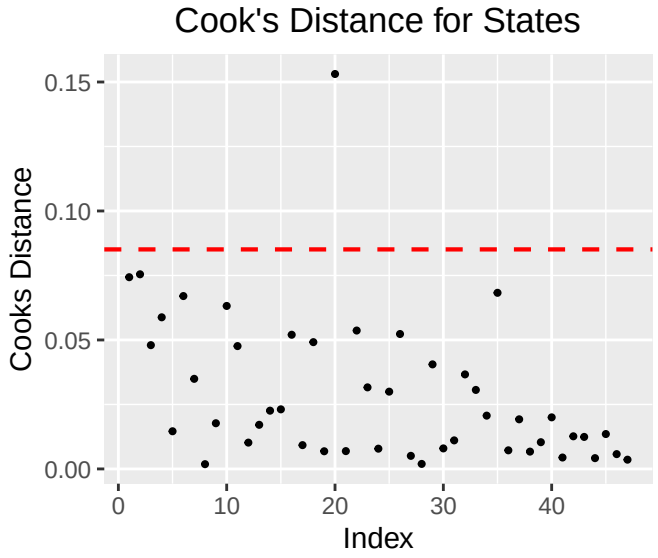
Table 8: Interaction

Model	LRT.p.value	BIC
without interaction		14751.51
+ quarter x bulk_purchase	0.1566	14772.17
+ quarter x mgstr2	0.4914	14820.71
+ bulk_purchase x mgstr2	0.1411	14771.93
+ quarter x source	0.0022	14831.35
+ bulk_purchase x source	0.5854	14783.18
+ source x mgstr2	0.9947	14948.55









Conclusion

Fixed Effects

Random Effects

Table 9: Estimates of fixed effects

	Estimate	exp(Estimate)	Std. Error	df	t value	Pr(>|t|)
(Intercept)	-0.6346	0.5301	0.0393	296.1819	-16.1290	0.0000
quarter2	0.0854	1.0891	0.0321	5550.5948	2.6578	0.0079
quarter3	0.0841	1.0877	0.0332	5555.0726	2.5349	0.0113
quarter4	0.0844	1.0881	0.0341	5551.1650	2.4759	0.0133
sourceHeard it	0.0633	1.0653	0.0335	5556.1197	1.8904	0.0588
sourceInternet	-0.0041	0.9959	0.0625	5555.5822	-0.0656	0.9477
sourceInternet Pharmacy	-0.3227	0.7242	0.1016	5548.9347	-3.1743	0.0015
sourcePersonal	-0.0398	0.9609	0.0281	5557.7473	-1.4157	0.1569
mgstr22 medium	-0.3816	0.6827	0.0279	5549.1951	-13.6631	0.0000
mgstr23 medium high	-0.7000	0.4966	0.0365	5554.7006	-19.1836	0.0000
mgstr24 high	-1.1197	0.3264	0.0420	5559.7107	-26.6889	0.0000
bulk_purchase1 Bulk purchase	-0.1141	0.8922	0.0296	5557.9880	-3.8488	0.0001

Table 10: Estimated random intercepts

grpvar	term	grp	condval	condsd	exp(condval)
state	(Intercept)	Massachusetts	0.2063	0.0761	1.2292
state	(Intercept)	New Jersey	0.1713	0.0784	1.1869
state	(Intercept)	Tennessee	0.1702	0.0611	1.1855
state	(Intercept)	Maine	0.1189	0.1091	1.1263
state	(Intercept)	Virginia	0.1144	0.0766	1.1212
state	(Intercept)	Mississippi	0.1130	0.0984	1.1197
state	(Intercept)	Louisiana	0.1127	0.0851	1.1193
state	(Intercept)	Maryland	0.0953	0.0784	1.1000
state	(Intercept)	West Virginia	0.0774	0.1037	1.0805
state	(Intercept)	Alaska	0.0762	0.1145	1.0792
state	(Intercept)	Kansas	0.0739	0.0898	1.0767
state	(Intercept)	Kentucky	0.0716	0.0843	1.0743
state	(Intercept)	Oklahoma	0.0712	0.0631	1.0738
state	(Intercept)	New Hampshire	0.0697	0.1154	1.0721
state	(Intercept)	Arkansas	0.0558	0.0804	1.0573
state	(Intercept)	Washington, DC	0.0546	0.1154	1.0561
state	(Intercept)	Utah	0.0244	0.0908	1.0247
state	(Intercept)	Nebraska	0.0221	0.0966	1.0224
state	(Intercept)	Alabama	0.0181	0.0682	1.0182
state	(Intercept)	Montana	0.0152	0.1083	1.0153
state	(Intercept)	South Dakota	0.0152	0.1185	1.0153
state	(Intercept)	Wisconsin	0.0120	0.0726	1.0121
state	(Intercept)	Iowa	0.0096	0.0839	1.0097
state	(Intercept)	Rhode Island	0.0074	0.1117	1.0074
state	(Intercept)	Wyoming	0.0021	0.1175	1.0021
state	(Intercept)	Florida	-0.0049	0.0415	0.9951
state	(Intercept)	Minnesota	-0.0054	0.0867	0.9946
state	(Intercept)	Indiana	-0.0109	0.0655	0.9891
state	(Intercept)	Washington	-0.0119	0.0550	0.9881
state	(Intercept)	North Dakota	-0.0139	0.1207	0.9862
state	(Intercept)	Missouri	-0.0150	0.0657	0.9851

state	(Intercept)	Colorado	-0.0157	0.0677	0.9844
state	(Intercept)	Oregon	-0.0181	0.0615	0.9820
state	(Intercept)	Pennsylvania	-0.0185	0.0523	0.9816
state	(Intercept)	Idaho	-0.0188	0.1051	0.9814
state	(Intercept)	Texas	-0.0342	0.0448	0.9663
state	(Intercept)	New York	-0.0385	0.0564	0.9623
state	(Intercept)	Ohio	-0.0412	0.0573	0.9596
state	(Intercept)	North Carolina	-0.0444	0.0668	0.9565
state	(Intercept)	Delaware	-0.0548	0.1108	0.9466
state	(Intercept)	New Mexico	-0.0792	0.0984	0.9239
state	(Intercept)	Illinois	-0.0917	0.0666	0.9124
state	(Intercept)	Hawaii	-0.0944	0.1009	0.9100
state	(Intercept)	Nevada	-0.1057	0.0664	0.8997
state	(Intercept)	Georgia	-0.1287	0.0629	0.8793
state	(Intercept)	Connecticut	-0.1363	0.0960	0.8726
state	(Intercept)	South Carolina	-0.1379	0.0775	0.8712
state	(Intercept)	California	-0.2098	0.0313	0.8107
state	(Intercept)	Arizona	-0.2192	0.0514	0.8031
state	(Intercept)	Michigan	-0.2297	0.0490	0.7948

Codes

```
knitr::opts_chunk$set(warning=FALSE, message = FALSE, cache = TRUE)
library(tidyverse)
library(janitor)
library(gridExtra)
library(kableExtra)
library(cowplot)
library(knitr)
library(magrittr)
library(dplyr)
library(readr)
library(tidyr)
library(broom)
library(lme4)
library(glmmTMB)
library(sjPlot)
library(coda)
library(naniar)
library(olsrr)
library(lmerTest)
library(lattice)
```

```
load('streetrx.RData')
```

```
na_check <- streetrx %>%
  filter(api_temp == 'morphine') %>%
  mutate_all( list( ~na_if(., '') ) ) %>%
  droplevels()

dim(na_check)
sum(is.na(na_check))

sum(is.na(na_check$Primary_Reason))

sum(is.na(na_check$source))

sum(is.na(na_check))

gg_miss_upset(na_check)
```

```
# subset for group drug

morph_data <- streetrx %>%
  filter(api_temp == 'morphine')

morph_data$Primary_Reason <- droplevels(morph_data$Primary_Reason)
levels(morph_data$Primary_Reason)[1] <- "8 Prefer not to answer"
levels(morph_data$Primary_Reason)[2] <- "8 Prefer not to answer"
```

```r
morph_data$source <- droplevels(morph_data$source)
levels(morph_data$source)[1] <- "Blank"

morph_data <- morph_data %>%
  filter(between(ppm, 0.000001, 10)) %>%
  mutate_all( list( ~na_if(., '') ) ) %>%
  drop_na() %>%
  clean_names() %>%
  mutate(
    quarter=substring(yq_pdate, 5, 5),
    year=substring(yq_pdate, 1, 4),
    state=recode_factor(droplevels(state), 'USA'='Unknown')
  )

nrow(morph_data)

sum(morph_data$ppm <=0)
```

```r
# remove extreme outliers based on quantiles

# untransformed density
p1 <- morph_data %>%
  ggplot(aes(x=ppm)) +
    geom_histogram(
      aes(y=..density..),
      color='black',
      linetype='dashed',
      size=0.5,
      fill='lightblue',
      alpha=0.5,
      bins=20
    ) +
    geom_density(size=0.75, bw=0.3) +
    ggtitle("Figure 1: Distribution of morphine ppm") +
    theme(plot.title = element_text(hjust = 0.5))



# log-transformed density
p2 <- morph_data %>%
  ggplot(aes(x=log(ppm))) +
    geom_histogram(
      aes(y=..density..),
      color='black',
      linetype='dashed',
      size=0.5,
      fill='lightcoral',
      alpha=0.5,
      bins=20
    ) +
    geom_density(size=0.75, bw=0.3) +
```

```
    xlim(-7, 3) +
    ggtitle("Figure 2: Distribution of morphine log(ppm)") +
    theme(plot.title = element_text(hjust = 0.5))

grid.arrange(p1, p2, ncol=2)
```

```
length(unique(morph_data$city))
length(unique(morph_data$state))
length(unique(morph_data$usa_region))
```

```
state_size <- morph_data %>%
  group_by(state) %>%
  summarise(n = n(), .groups = "drop")  %>%
  arrange(n) %>%
  pivot_wider(
    names_from=state,
    values_from=n
  )

state_size %>%
  dplyr::select(1:5) %>%
  kable(
    caption = '5 States with Smallest Sample Size',
    align='c',
    booktabs=TRUE) %>%
  kable_styling(latex_options = c('hold_position'))
```

```
# remove low sample size states
morph_data <- morph_data %>%
  mutate(state=as.character(state)) %>%
  filter(!state %in% c(
    'Puerto Rico', 'Vermont'
  ))
```

```
morph_state <- morph_data %>%
  filter(state %in% state.name) %>%
  mutate(state_abv=state.abb[match(state,state.name)])


grand_mean <- mean(morph_state$ppm)

p3 <- morph_state %>%
  group_by(state_abv) %>%
  summarise(n = n(), mean = mean(ppm)) %>%
  ggplot(aes(x=n, y=mean)) +
    geom_hline(
      aes(yintercept=grand_mean),
      linetype='dashed',
      color='red',
      size=0.75
```

```
    ) +
    geom_point() +
    labs(x='sample size', y='mean log(ppm)') +
    ggtitle("Figure 3: Group mean vs. sample size") +
    theme(plot.title = element_text(hjust = 0.5))


p4 <- morph_state %>%
  ggplot(aes(y=log(ppm), x=state_abv)) +
    geom_boxplot(
      fill=rainbow(49),
      alpha=0.5
    ) +
    scale_x_discrete(guide=guide_axis(angle = 90)) +
    ggtitle("Figure 4: Boxplot of log(pmm) across states") +
    theme(plot.title = element_text(hjust = 0.5)) +
    labs(x='state')

cowplot::plot_grid(p3, p4, rel_widths = c(1, 2))
```

```
t1 <- morph_state %>%
  group_by(usa_region) %>%
  summarise(n=n(), mean=round(mean(log(ppm)), 3)) %>%
  tableGrob()

p5 <- morph_state %>%
  ggplot(aes(y=log(ppm), x=usa_region)) +
    geom_boxplot() +
    stat_summary(
      fun.y=mean,
      geom='point',
      color='red',
      size=3
    ) +
    ggtitle("Figure 5: Boxplot of log(pmm) across regions") +
    theme(plot.title = element_text(hjust = 0.5))

grid.arrange(t1, p5, ncol=2, widths=c(2, 2))
```

```
min(as.Date(morph_data$price_date, "%m/%d/%y")) #2013-01-01
morph_data %>% group_by(year) %>% summarise(n = n())
```

```
# remove data prior to 2010
morph_data <- morph_data %>%
  mutate(Year=as.character(year)) %>%
  filter(!year %in% c(
    1969, 2000, 2002, 2005
  ))
```

```r
# date_diff
morph_data <- morph_data %>%
  mutate(date_diff = as.numeric(
    as.Date(morph_data$price_date, "%m/%d/%y") - as.Date("2010-01-01")
    )
  )
```

```r
morph_data %>%
  ggplot(aes(x=date_diff)) +
    geom_histogram(
      aes(y=..density..),
      color='black',
      linetype='dashed',
      size=0.5,
      fill='lightblue',
      alpha=0.5,
      bins=30
    ) +
    geom_density(size=0.75, bw=100) +
  ggtitle("Date Distribution") +
  theme(plot.title = element_text(hjust = 0.5))
```

```r
morph_data %>%
  ggplot(aes(x=date_diff, y=log(ppm))) +
    geom_point() +
    geom_smooth() +
    theme_bw()
```

```r
yearplot <- morph_data %>%
  ggplot(aes(x = year,y = log(ppm))) +
  geom_boxplot() +
  labs(x='Year') +
  stat_summary(
    fun.y=mean,
    geom='point',
    color='red',
    size=3
  ) +
  ggtitle("Figure 6: log(pmm) vs. years") +
  theme(plot.title = element_text(hjust = 0.5))



quarterplot <- morph_data %>%
  ggplot(aes(x = quarter,y = log(ppm))) +
  geom_boxplot() +
  labs(
    x='Quarter',
    y=''
  )  +
  stat_summary(
```

```
    fun.y=mean,
    geom='point',
    color='red',
    size=3
  ) +
  ggtitle("Figure 7: log(pmm) vs. quarters") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(yearplot, quarterplot, ncol=2)
```

```
plot1 <- morph_data %>%
  ggplot(aes(x = bulk_purchase,y = log(ppm))) +
  geom_boxplot() +
  labs(x='Bulk Purchase') +
  stat_summary(
    fun.y=mean,
    geom='point',
    color='red',
    size=3
  ) +
  ggtitle("Figure 8: log(pmm) vs. bulk purchase") +
  theme(plot.title = element_text(hjust = 0.5))


# unique(morph_data$source)

# combine internet levels into single level
morph_data <- morph_data %>%
  mutate(source=replace(
    source, !source %in% c(
      "Blank",
      'Personal',
      'Heard it',
      'Internet',
      'Internet Pharmacy',
      'Drug forum'
    ), 'Internet'
  )) %>%
  droplevels()


morph_data <- morph_data %>%
  mutate(source=as.character(source)) %>%
  filter(source != "Drug forum")

plot2 <- morph_data %>%
  ggplot(aes(x = source,y = log(ppm))) +
  geom_boxplot() +
  stat_summary(
    fun.y=mean,
    geom='point',
```

```
    color='red',
    size=2
  ) +
  ggtitle("Figure 9: log(pmm) vs. source") +
  theme(plot.title = element_text(hjust = 0.5))


grid.arrange(plot1, plot2, ncol =2)
```

```
morph_data %>%
  ggplot(aes(x=mgstr, y=log(ppm))) +
    geom_point() +
    geom_smooth() +
    theme_bw()

# morph_data %>%
#   ggplot(aes(x=log(mgstr), y=log(ppm))) +
#     geom_point() +
#     geom_smooth() +
#     theme_bw()

morph_data %>%
  ggplot(aes(x=mgstr)) +
    geom_histogram(
      aes(y=..density..),
      color='black',
      linetype='dashed',
      size=0.5,
      fill='lightblue',
      alpha=0.5,
      bins=10
    ) +
    geom_density(size=0.75, bw=7.5) +
    labs(title='mgstr Distribution') +
  labs(title='log(pmm) vs. sources') +
    theme_bw()
```

```
# check for random slopes
morph_data %>% ggplot(aes(x = mgstr, y = log(ppm))) +
  geom_point() +
  geom_smooth() +
  theme_bw() +
  facet_wrap('usa_region', scales = "fixed")
```

```
morph_data %>%
  group_by(mgstr) %>%
  summarize(n = n()) %>%
  pivot_wider(
    names_from=mgstr,
    values_from=n
  ) %>%
```

```
  kable(
    caption='Sample Size for mgstr Levels',
    align='c',
    booktabs=TRUE
  ) %>%
  kable_styling(latex_options = c('hold_position'))

# inspect mgstr value quantiles
quantile(morph_data$mgstr, c(0.25, 0.5, 0.75)) %>%
  data.frame() %>%
  rename('mgstr'='.') %>%
  kable()
```

```
## here we decide to re-code mgstr by quantile
morph_data <- morph_data %>%
  mutate(mgstr2 = case_when(
    mgstr <= 15               ~ "1 low",
    mgstr >15 & mgstr <= 30   ~ "2 medium",
    mgstr >30 & mgstr <= 60   ~ "3 medium high",
    mgstr > 60             ~ "4 high")
  )

plot1 <- morph_data %>%
  ggplot(aes(x=mgstr2 ,y=log(ppm))) +
  geom_boxplot() +
  labs(y="log(ppm)", x="Strength")  +
  stat_summary(
    fun.y=mean,
    geom='point',
    color='red',
    size=3
  ) +
  ggtitle("log(pmm) vs. strength") +
  theme(plot.title = element_text(hjust = 0.5))


plot2 <- morph_data %>%
  ggplot(aes(x = primary_reason,y =log(ppm))) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "log(ppm)", y = "Reason")  +
    stat_summary(
      fun.y=mean,
      geom='point',
      color='red',
      size=3
    ) +
  ggtitle("log(pmm) vs. reasons") +
  theme(plot.title = element_text(hjust = 0.5))

grid.arrange(arrangeGrob(plot1, ncol=1, nrow=1),
```

```
             arrangeGrob(plot2, ncol=1, nrow=1), widths=c(1,2))
```

Interaction Plot

```
morph_data %>%
  ggplot(aes(x = mgstr2,y =log(ppm), fill = bulk_purchase)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "log(ppm)", y = "Strength")  +
  ggtitle("log(pmm) vs. bulk purchase x strength") +
  theme(plot.title = element_text(hjust = 0.5))

morph_data %>%
  ggplot(aes(x = mgstr2,y =log(ppm), fill = quarter)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "log(ppm)", y = "Strength")  +
  ggtitle("log(pmm) vs. quarter x strength") +
  theme(plot.title = element_text(hjust = 0.5))

morph_data %>%
  ggplot(aes(x = primary_reason,y =log(ppm), fill = bulk_purchase)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "Primary Reason", y = "log(ppm)")  +
  ggtitle("log(pmm) vs. bulk purchase x primary reason") +
  theme(plot.title = element_text(hjust = 0.5))

morph_data %>%
  ggplot(aes(x = primary_reason,y =log(ppm), fill = quarter)) +
  geom_boxplot() +
  coord_flip() +
  labs(x = "Primary Reason", y = "log(ppm)")  +
  ggtitle("log(pmm) vs. quarter x primary reason") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
# group by city
mod_1 <- lmer(data=morph_data, log(ppm) ~ (1 |city), REML=F)

# group by state
mod_2 <- lmer(data=morph_data, log(ppm) ~ (1 |state), REML=F)

# group by region
mod_3 <- lmer(data=morph_data, log(ppm) ~ (1 |usa_region), REML=F)

aic_score <- sapply(c(mod_1, mod_2, mod_3), AIC)
bic_score <- sapply(c(mod_1, mod_2, mod_3), BIC)

data.frame('Grouping' = c('City', 'State', 'Region'), 'AIC' = aic_score,
```

```
           'BIC' = bic_score) %>%
  kable(caption = "AIC and BIC for different grouping variables")
```

```
# appendix
# group by city
mod_full_1 <- lmer(data=morph_data, log(ppm) ~ (1 |city) + date_diff + quarter
                   + year + mgstr2 +
                 bulk_purchase + primary_reason + source, REML=F)

# group by state
mod_full_2 <- lmer(data=morph_data, log(ppm) ~ (1 |state) + date_diff + quarter
                   + year + mgstr2 +
                 bulk_purchase + primary_reason + source, REML=F)

# group by region
mod_full_3 <- lmer(data=morph_data, log(ppm) ~ (1 |usa_region) + date_diff +
                     quarter + year + mgstr2 +
                 bulk_purchase + primary_reason + source, REML=F)

aic_score <- sapply(c(mod_full_1, mod_full_2, mod_full_3), AIC)
bic_score <- sapply(c(mod_full_1, mod_full_2, mod_full_3), BIC)

data.frame('Grouping' = c('City', 'State', 'Region'), 'AIC' = aic_score,
           'BIC' = bic_score) %>%
  kable() %>%
    kable_styling(latex_options = c("hold_position","striped"))
```

```
modelAll <- lmer(log(ppm) ~ quarter + primary_reason + mgstr2 + bulk_purchase
                 + year + date_diff + source + (1|state), data = morph_data,
                 REML=F)
step(modelAll)
```

```
model <- c("(1|state)",
           "(1|state) + mgstr2",
           "(1|state) + mgstr2 + bulk_purchase",
           "(1|state) + mgstr2 + bulk_purchase + year",
           "(1|state) + mgstr2 + bulk_purchase + quarter",
           "(1|state) + mgstr2 + bulk_purchase + date_diff",
           "(1|state) + mgstr2 + bulk_purchase + quarter + source",
           "(1|state) + mgstr2 + bulk_purchase + quarter + source +
           primary_reason")
LRT <- c("",
         round(anova(modela,modelb)$`Pr(>Chisq)`[2],4),
         round(anova(modelb,modelc)$`Pr(>Chisq)`[2],4),
         round(anova(modelc,modeld)$`Pr(>Chisq)`[2],4),
         round(anova(modelc,modele)$`Pr(>Chisq)`[2],4),
         round(anova(modelc,modelf)$`Pr(>Chisq)`[2],4),
         round(anova(modele,modelg)$`Pr(>Chisq)`[2],4),
         round(anova(modelg,modelh)$`Pr(>Chisq)`[2],4))
BIC_score1 <- sapply(c(modela, modelb, modelc, modeld, modele, modelf, modelg,
                       modelh), BIC)
```

```
data.frame("Model" = model, 'LRT p-value' = LRT, 'BIC' = BIC_score1) %>%
  kable(caption = "Forward model selection") %>%
    kable_styling(latex_options = c("hold_position","striped"))
```

```
modelg <- lmer(log(ppm) ~  quarter + source + mgstr2 + bulk_purchase +
                 (1|state), data = morph_data, REML=F)
```

```
plot_qq <- function(model) {
  df <- data.frame(
    res=residuals(model, scaled=TRUE)
  )

  p <- ggplot(df, aes(sample=res)) +
    stat_qq(
      size=0.75
    ) +
    stat_qq_line(
      linetype='dashed',
      color='red',
      size=0.5
    ) +
    labs(
      x='Theoretical Quantiles',
      y='Standardized Residuals'
    ) +
    ggtitle("Normal QQ for Residuals") +
    theme(plot.title = element_text(hjust = 0.5))

  return(p)
}


plot_ranef_qq <- function(model) {
  df <- data.frame(
    res=ranef(model)[[1]][[1]]
  )

  p <- ggplot(df, aes(sample=res)) +
    stat_qq(
      size=0.5
    ) +
    stat_qq_line(
      linetype='dashed',
      color='red',
      size=0.5
    ) +
    labs(
      x='Theoretical Quantiles',
      y='State Intercept'
    ) +
    ggtitle("Normal QQ for Random Effects") +
```

```
  theme(plot.title = element_text(hjust = 0.5))

  return(p)
}

plot_res_fit <- function(model) {
  df <- data.frame(
    res=residuals(model),
    fit=fitted(model)
  )

  p <- ggplot(df, aes(x=fit, y=res)) +
    geom_point(
      size=0.75
    ) +
    geom_hline(
      yintercept=0,
      linetype="dashed"
    ) +
    geom_smooth() +
    labs(
      x='Fitted',
      y='Residuals'
    ) +
    ggtitle("Residuals vs. Fitted") +
  theme(plot.title = element_text(hjust = 0.5))

  return(p)
}

plot_scale_loc <- function(model) {
  df <- data.frame(
    res=sqrt(residuals(model, scaled=TRUE)),
    fit=fitted(model)
  )

  p <- ggplot(df, aes(x=fit, y=res)) +
    geom_point(
      size=0.75
    ) +
    geom_smooth() +
    labs(
      x='Fitted',
      y=expression(sqrt('Standardized Residuals'))
    ) +
    ggtitle("Scale-Location") +
  theme(plot.title = element_text(hjust = 0.5))

  return(p)
}
```

```r
plot_res_dens <- function(model) {
  df <- data.frame(
    res=residuals(model)
  )

  p <- ggplot(df, aes(x=res)) +
    geom_density() +
    labs(
      x='Residuals',
      y='Density'
    ) +
    ggtitle("Residuals Density") +
  theme(plot.title = element_text(hjust = 0.5))

  return(p)
}

plot_cooks_distance <- function(model1){
  model_inf<- influence(model1, group = "state")
  data <- model.frame(model1)
  cooks_distance <- cooks.distance(model_inf)
  cutline <- 4 / length(unique(data$state))
  infindiv <- cooks_distance > cutline

  p <- ggplot(data=NULL, aes(x=1:length(unique(data$state)), y=cooks_distance)) +
    geom_point(
      size=0.75
    ) +
    geom_hline(
      yintercept=cutline,
      linetype='dashed',
      color='red',
      size=0.75
    ) +
    labs(
      x='Index',
      y='Cooks Distance'
    ) +
  ggtitle("Cook's Distance for States") +
  theme(plot.title = element_text(hjust = 0.5))
  return(p)
}

model_diag <- function(model) {
  p1 <- plot_res_fit(model)
  p2 <- plot_qq(model)
  p3 <- plot_ranef_qq(model)
  p4 <- plot_cooks_distance(model)

  cowplot::plot_grid(p1, p2, p3, p4, nrow=2)
}
```

```
model_diag(modelg)
```

```
# remove lowest residual data point
morph_data2 <- morph_data[-which.min(resid(modelg)),]


model_g2 <- lmer(log(ppm) ~  quarter + source + mgstr2 + bulk_purchase +
                   (1|state), data = morph_data2, REML=F)


# view_coef(model_g2)
# view_params(model_g2)
# model_diag(model_g2)
```

```
model_g2_inf<- influence(model_g2, group = "state")
```

```
cooks_distance <- cooks.distance(model_g2_inf)
cutline <- 4 / length(unique(morph_data2$state))
infindiv <- cooks_distance > cutline

ggplot(data=NULL, aes(x=1:length(unique(morph_data2$state)), y=cooks_distance)) +
  geom_point() +
  geom_hline(
    yintercept=cutline,
    linetype='dashed',
    color='red',
    size=0.75
  ) +
  labs(
    x='Index',
    y='Cooks Distance'
  ) +
  theme_bw()


data.frame(
  rownames(model_g2_inf$`fixed.effects[-state]`),
  round(cooks_distance, 4),
  infindiv
) %>%
  filter(infindiv == TRUE) %>%
  dplyr::select(1:2) %>%
  rename(`State`=1, `Cook's Distance`=2) %>%
  kable() %>%
  kable_classic(full_width=FALSE)
```

```
# remove three most influential states
morph_data3 <- morph_data2 %>%
  filter(!state %in% c('Florida', 'California','Pennsylvania'))
```

```
model_g3 <- lmer(log(ppm) ~  quarter + source + mgstr2 + bulk_purchase +
                   (1|state), data = morph_data3, REML=F)

model_diag(model_g3)
```

```
# view coefficient estimates
view_coef <- function(model) {
  summary(model)$coefficients %>%
    as.data.frame() %>%
    mutate(`exp(Estimate)`=exp(Estimate)) %>%
    relocate(`exp(Estimate)`, .after=Estimate) %>%
    kable(caption = "Estimates of fixed effects",
      digits = 4) %>%
    kable_classic(full_width=FALSE)
}

# view parameter estimates
view_params <- function(model) {
  params <- summary(model)$varcor %>%
    as.data.frame() %>%
    dplyr::select(vcov)
  params <- t(params)
  rownames(params) <- c('Estimate')

  kable(params,
        caption = "Estimates of random effects",
        col.names = c('$\\tau^2$', '$\\sigma^2$'),
        digits = 4,
        format = 'latex',
        escape = FALSE) %>%
  kable_classic(full_width=FALSE)
}
```

```
view_coef(model_g2) %>%
  kable_styling(latex_options = c("hold_position","striped"))
```

```
view_params(model_g2) %>%
  kable_styling(latex_options = c("hold_position","striped"))
```

```
# view intercept estimates and intervals
dotplot(ranef(model_g2, condVar = TRUE))$state
```