

Case Study 1

Marc Brooks (Presenter), Bo Liu (Programmer), Shirley Mathur (Writer), Aasha Reddy (Checker and C

10/6/2021

10/17 Meeting, things to discuss - should we move the Cooks distance to the appendix? - discussion of qq plot is needed - should we interpret the Bayesian model? - Current bayesian model does not use the results from the frequentist model, does it? - Model building section - only include the ex_result_int table? - finish data cleaning section writing - finish model building section writing - Need to make a code appendix once the report is completely finalized

To do (10/15): - Data cleaning - get rid of states with 1 observation - EDA - Cooks distance by state for assumption checking - For qq plot, don't try other transformations (square is too hard to interpret). Log transform takes the range to the whole real line - modeling - include interactions using exhaustive search - include year as a factor - include empirical bayes (use BIC to select frequentist model then use frequentist results to inform priors, use uninformative variance priors)

To do: - [Aasha] get rid of states with 1 obs - [Bo] change exhaustive search function - [Marc] Cooks distance by state for assumption checking - [Bo] BRMS - [Shirley] interpret Bayesian - [Shirley] Data cleaning report writing - [Aasha] EDA Report Writing - [Bo] Modeling Report Writing - [Marc] Interpretation Report Writing

Introduction

Prescription opioid diversion and abuse are major public health issues, and street prices provide an indicator of drug availability, demand, and abuse potential. Using StreetRx data, we aim to investigate factors related to the price per mg of Methadone.

StreetRx (streetrx.com) is a web-based citizen reporting tool enabling real-time collection of street price data on diverted pharmaceutical substances. Based on principles of crowdsourcing 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 intelligence into an otherwise opaque black market, providing a novel data set for public health surveillance, specifically for controlled substances.

Our goal is to investigate factors related to the price per mg of Methadone, accounting for potential clustering by location and exploring heterogeneity in pricing by location. Our data contains the following factors, and we will explore how the factors in the dataset are or are not associated with pricing per milligram.

We first clean data and conduct exploratory data analysis (EDA) to assess any necessary transformations. We also conduct EDA to assess what type of model we should build, including which variables to include and to assess whether random intercepts or slopes would be helpful. We also use exhaustive search using BIC to perform variable selection. In our final model, we include random intercepts by state, and fixed effects for mgstr (), source (), and bulk_purchase (). We use estimates obtained from a frequentist version of the hierarchical model to inform priors for the final Bayesian version of the model. We find that _____.

Research Questions:

- Which variables are associated with pricing per milligram of Methadone?
- Is there heterogeneity in pricing of Methadone by location?

Data and Cleaning

We first examine missing data in the streetrx data. We substitute NA for all missing values. We see that some variables, such as ppm, city, source, mgstr and primary_reason have many missing values. For the purpose of our model, we will not use the primary_reason variable, as they contain the most missing observations. We also note that individuals self-report their city, state, and country, so there are some data entry errors. For instance, some observations report purchased in “New York” vs. “New York Manhattan” vs. “New York City”, which all refer to the same city. Thus, this variable may not be reliable as a grouping variable to explore heterogeneity within location. This is not an issue with State, so we choose to use state as our grouping variable. We also note that all purchases were made in the USA.

Then, we proceeded to modify variables and remove observation as needed. We modified the types of the variables as needed to make sure that variables were reflected as factors or numeric depending on the variable. We also noted that all of the observations with a missing value for price per milligram also were missing a value for the dosage, so we removed all of these observations as it is nonsensical to include observations that are missing a value for the response variable. The fact that these variables also were missing the dosage value then leads to us having no missing dosage values in the data we use for modeling, which is helpful. Following that, from our EDA we noticed that most of the dosage values were either 5, 10, or 40 mg, so we decided to remove the few observations that did not have these values (these were less than 10 observations did not have these dosage), and then we coded the dosage variable (mgstr) as a factor variable since it only has three levels. Following that, we noted while examining the that two of the observations had stated 1969 as the year, which did not make sense considering StreetRx did not exist at the time, so we removed these observations as well. Since we decided to use stats as our grouping variable, we also removed observations that had input “USA” as the state, as these would not actually give us information about heterogeneity in price per milligram values amongst states. We also combined all of the various websites that were reported as sources and coded them all as “Internet” since all of these sites are essentially the same as having sourced from the Internet, and that greatly reduced the number of levels we would have for the Source variable. Finally, we removed any states with only one observation as it would not make sense to try to estimate within-group variance for states with only one observation. However, this only removed two states/territories, Washington D.C. and Guam, so we can still get useful insights into the heterogeneity present in prices per milligram across all of the other states and territories in the dataset.

Table 1: Variable Descriptions

Variable	Description
ppm	Price per mg (outcome of interest)
yq_pdate	Year and quarter drug was purchased
price_date	Date of the reported purchase
city	city purchased
state	state purchased
country	country purchased
USA_region	northeast, midwest, west, south, or other/unknown
source	source of information
api_temp	active ingredient of drug of interest, in our case Methadone)
form_temp	formulation of the drug (e.g., pill, patch, suppository)
mgstr	dosage strength in mg of the units purchased
bulk_purchase	indicator for purchase of 10+ units at once
Primary_Reason	primary reason for purchase

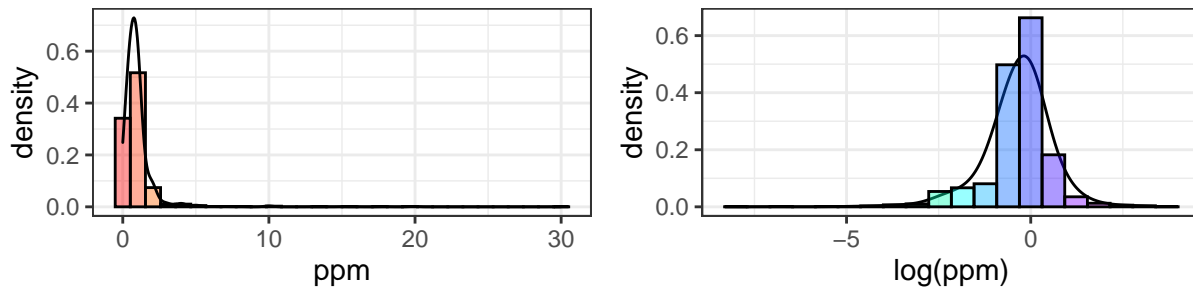
EDA

Checking distribution of outcome variable (ppm)

We explore the data first by examining the distribution of our outcome variable, price per mg (ppm). We next examine the distribution of our outcome variable, ppm (price per mg). We can see that ppm spans a large range of values, from \$0.00025/mg to \$40/mg. We note from the below histogram that the distribution is highly right-skewed. To satisfy the conditional distribution assumption, we want to aim for ppm to be normally distributed and symmetric. This of course only examines the marginal distribution of ppm, but the idea is that this may carry over into the conditional distribution, which we examine after the model fitting process

We choose to do a log transformation of ppm. We can see in the below plot that the histogram of $\log(\text{ppm})$ is relatively normally distributed and symmetric. Using a log transformation is a good choice as well because our raw data ppm observation are all greater than 0, and log transformations are still interpretable, which is important in this case study.

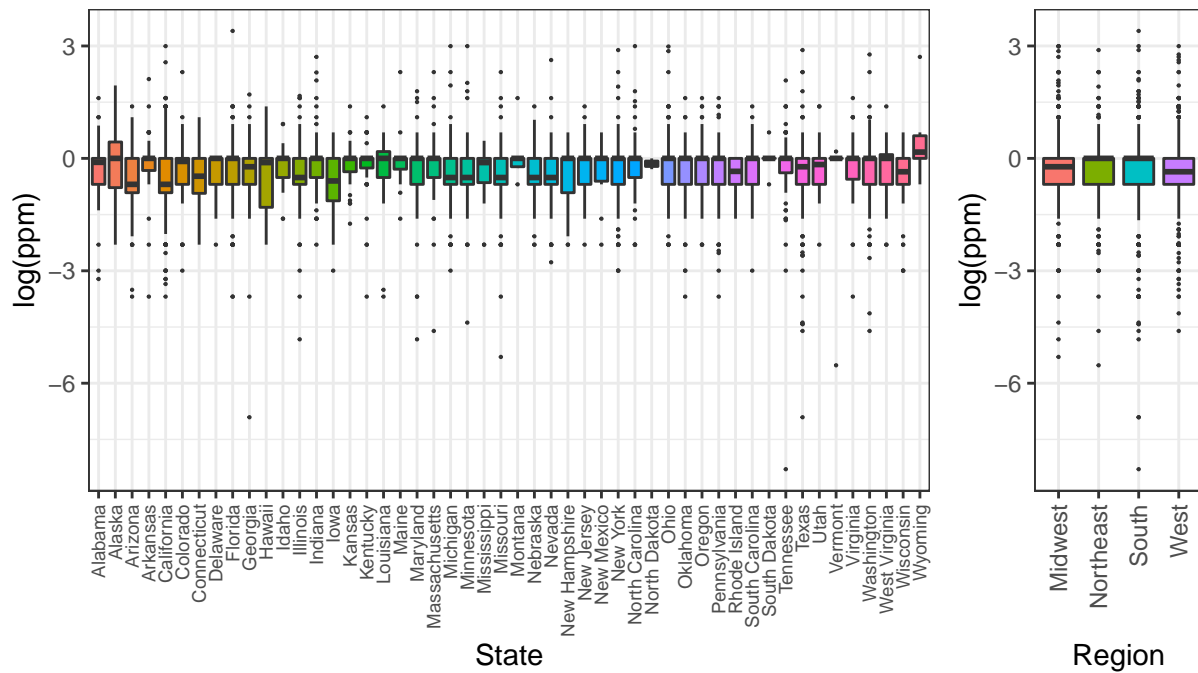
Distribution of ppm and $\log(\text{ppm})$



Assessing random intercepts

There are a few options for grouping variables for a random intercept. We exclude `city` as we noted previously that this field is highly erroneous. We assess both `state` and `region` as potential grouping variables by examining heterogeneity of $\log(\text{ppm})$ among both states and regions. From the below plots, we can see that there is not much variation of $\log(\text{ppm})$ by region, but there is some variation by state. Thus, we will include a random intercept by state in our model.

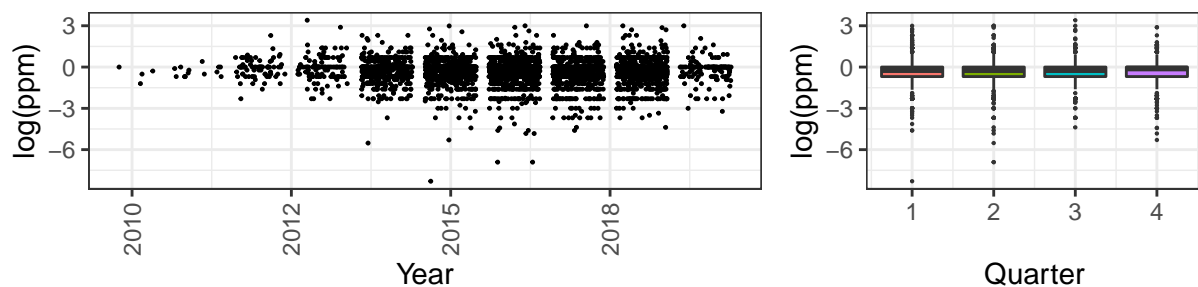
Log(ppm) by state and region



Assessing relationship of variables with log(ppm)

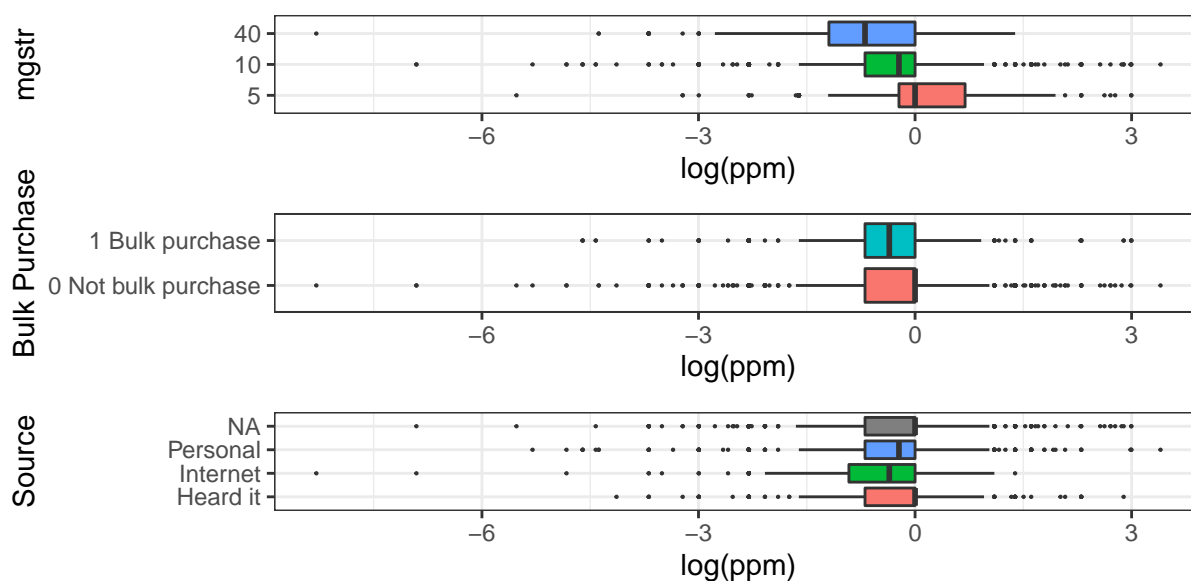
We next assess relationships of variables in our dataset with log(ppm). This is useful to understand which variables may be most helpful to include as fixed effects in our model. We first examine year vs. log(ppm) and see that there is some evidence of a relationship. We also feel it is important to test for effects of year on log(ppm) to account for any potential inflation in the price of Methadone. Thus, we choose to include year as a fixed effect in our model selection process. We also note from the boxplot below that quarter does not seem to have variation by log(ppm), so we do not include this in our model. We wanted to consider this variable to account for any potential seasonality in the price of Methadone.

Log(ppm) by year and quarter



We next assess the relationship of mgstr, bulk_purchase, and source by log(ppm) in the below boxplots. We see that all of these variables seem to have differences in log(ppm) by their respective levels, thus we choose to include mgstr, bulk_purchase, and source as fixed effects in our model selection process.

Log(ppm) vs. mgstr, bulk purchase, and source



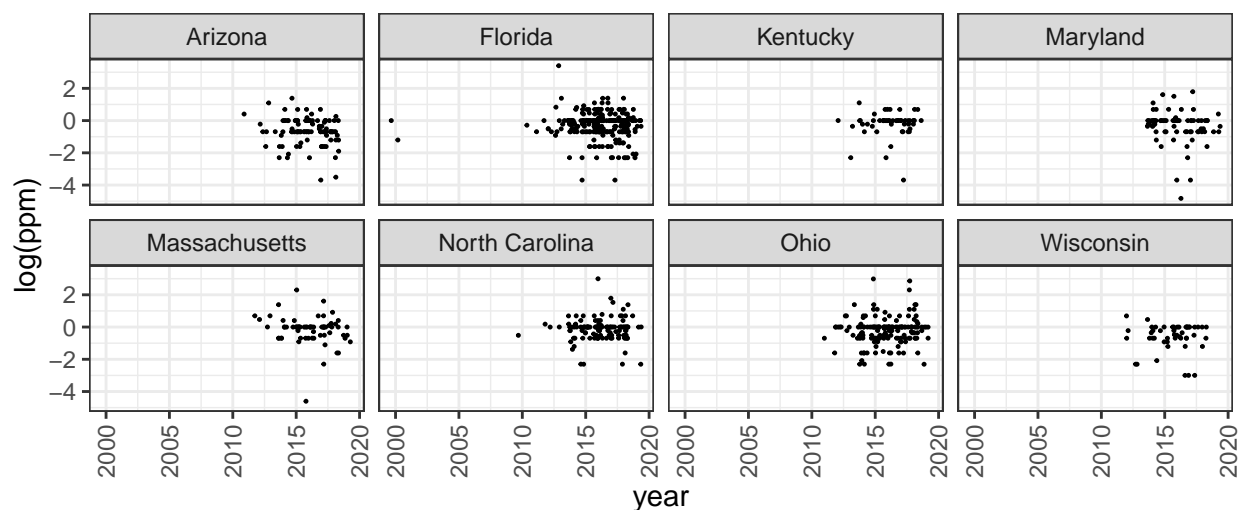
We note that `form_temp` (formulation of the drug as pill, patch, etc.) is always pill for Methadone, so we do not consider it as a potential variable in our model. The last variable in our dataset is primary reason for the purchase of Methadone. We note in the data cleaning section that this variable contains many missing observations, so we choose to exclude it.

Assessing Random Slopes

We next assess whether random slopes of our chosen variables by state would be useful. We note that we only have one continuous variable, year. In the below plots, we do not examine the trend of variables by all states. Instead, we filter for states with larger than 30 observations, and then choose a random sample of 8 states.

We first examine the trend of the relationship of year with $\log(\text{ppm})$ across 8 random states. We see that there is no distinguishable difference so we choose not to include a random slope of year by state.

Distribution of $\log(\text{ppm})$ by year and 8 random states



We then examine differences in the levels of `mgstr`, `source`, and `bulk_purchase` vs. $\log(\text{ppm})$ by state. We created boxplots of each level of `mgstr`, `source`, and `bulk_purchase` vs. $\log(\text{ppm})$ by state and did not find

evidence of any major difference between the levels of the factors vs. $\log(\text{ppm})$ by state. Thus we choose not to include any random slopes in our model. The boxplots can be found in the appendix.

Interactions

Next we assess whether any interactions would be useful to include in our model. In our EDA, we examined plots of all 2-way interactions even though we do not include plots here. Outside of `mgstr` and `quarter`, `bulk_purchase` and `quarter`, `source` and `mgstr`, and `bulk_purchase` and `mgstr` there was not strong evidence for other interaction effects. Even for those listed above the evidence was not substantial in our EDA, and some of the variation is likely due to a lack of observations for certain interaction terms. However, we test for inclusion of all 2-way interactions in our exhaustive search using BIC to make sure we capture any important interactions.

Overall choices and next steps

Through EDA, we have made the decision to include a random intercept by state in our model selection process, as well as fixed effects for year, `mgstr`, `bulk_purchase` and `source`. We will proceed with using BIC using exhaustive search to choose the best combination of fixed effects and 2-way interactions.

```
<<<<<<< HEAD state_rand <- as.character(state_rand)

streetrx %>% filter(state %in% c(state_rand)) %>% ggplot(., aes(x = source, y = log(ppm))) +
  geom_boxplot() + facet_wrap(~state) + labs(title = "Distribution of log(ppm) by source and 8 random
states")
```

3) year and state

`\begin{center}\includegraphics{case_study_01_report_files/figure-latex/unnamed-chunk-12-1} \end{center}`

4) bulk_purchase and state

`\begin{center}\includegraphics{case_study_01_report_files/figure-latex/unnamed-chunk-13-1} \end{center}`

5. Model

=====

Model Building and Selection

```
>>>>>>> main
```

****I do not have much information about what the best model should be, given that we are still in early**

1) Grouping variable

- ``USA_region`` and ``state`` are natural grouping variables - and required for the analysis purpose.
- Grouping on binary variables is equivalent to a random slope for it. If we decide to put a random slope
- Other categorical variables are subject on how many levels we keep.

2) Base model

State: $\$s\$$.

$\$y_{\{is\}} = \mu + \alpha_{\{s\}} + \epsilon_{\{is\}}.\$$

Where $y_{i,s}$ is the ppm for purchase i in state s

See how much heterogeneity.

*May compare the fitted variance of regional level means and state level means. If regional level means

AIC BIC logLik deviance df.resid

10568 10587 -5281 10562 4158

3) Adding predictors

Criterion: BIC(?) for Bayesian predictor selection.

We do an exhaustive search.

```
\begin{table}

\caption{\label{tab:unnamed-chunk-18}Exhaustive search of fixed effects using BIC}
\centering
\begin{tabular}{t}{l|r}
\hline
model & BIC\\
\hline
log(ppm) \textasciitilde{ } mgstr + bulk\_purchase + source + (1 | state) & 6413\\
\hline
log(ppm) \textasciitilde{ } mgstr + source + (1 | state) & 6418\\
\hline
log(ppm) \textasciitilde{ } year + mgstr + bulk\_purchase + source + (1 | state) & 6420\\
\hline
log(ppm) \textasciitilde{ } year + mgstr + source + (1 | state) & 6426\\
\hline
log(ppm) \textasciitilde{ } bulk\_purchase + source + (1 | state) & 6558\\
\hline
log(ppm) \textasciitilde{ } source + (1 | state) & 6560\\
\hline
log(ppm) \textasciitilde{ } year + bulk\_purchase + source + (1 | state) & 6562\\
\hline
log(ppm) \textasciitilde{ } year + source + (1 | state) & 6565\\
\hline
log(ppm) \textasciitilde{ } mgstr + bulk\_purchase + (1 | state) & 10347\\
\hline
log(ppm) \textasciitilde{ } mgstr + (1 | state) & 10355\end{tabular}
```

```

\hline
log(ppm) \textasciitilde{} year + mgstr + bulk\_purchase + (1 | state) & 10355\\
\hline
log(ppm) \textasciitilde{} year + mgstr + (1 | state) & 10363\\
\hline
log(ppm) \textasciitilde{} bulk\_purchase + (1 | state) & 10582\\
\hline
log(ppm) \textasciitilde{} year + bulk\_purchase + (1 | state) & 10587\\
\hline
log(ppm) \textasciitilde{} 1 + (1 | state) & 10587\\
\hline
log(ppm) \textasciitilde{} year + (1 | state) & 10592\\
\hline
\end{tabular}
\end{table}

\begin{table}

\caption{\label{tab:unnamed-chunk-19}Exhaustive search of fixed effects using BIC}
\centering
\begin{tabular}[t]{l|r}
\hline
model & BIC\\
\hline
log(ppm) \textasciitilde{} mgstr + bulk\_purchase + source + (1 | state) & 6413\\
\hline
log(ppm) \textasciitilde{} mgstr + bulk\_purchase:source + (1 | state) & 6418\\
\hline
log(ppm) \textasciitilde{} mgstr + source + bulk\_purchase:source + (1 | state) & 6418\\
\hline
log(ppm) \textasciitilde{} mgstr + bulk\_purchase + source + bulk\_purchase:source + (1 | state) & 6418\\
\hline
log(ppm) \textasciitilde{} mgstr + bulk\_purchase + bulk\_purchase:source + (1 | state) & 6418\\
\hline
log(ppm) \textasciitilde{} mgstr + source + (1 | state) & 6418\\
\hline
log(ppm) \textasciitilde{} mgstr + year:bulk\_purchase + year:source + (1 | state) & 6420\\
\hline
log(ppm) \textasciitilde{} year + mgstr + year:bulk\_purchase + year:source + (1 | state) & 6420\\
\hline
log(ppm) \textasciitilde{} year + mgstr + source + year:bulk\_purchase + (1 | state) & 6420\\
\hline
log(ppm) \textasciitilde{} mgstr + source + year:bulk\_purchase + (1 | state) & 6420\\
\hline
\end{tabular}
\end{table}

```

Our best model is

\$\$

```

\begin{aligned}
y_{is} = & \mu + \alpha_{is} + \beta_1 I(mgstr_{is} = 10) + \beta_2 I(mgstr_{is} = 40) + \beta_3 I(bulkp_{is} = 1) \\
& + \beta_4 I(source_{is} = internet) + \beta_4 I(source_{is} = personal) + \epsilon_{is}
\end{aligned}

```


\end{aligned}

$$$$

$$\begin{aligned}
&\alpha_s \sim \text{Normal}(0, \tau^2) \\
&\epsilon_{is} \sim \text{Normal}(0, \sigma^2)
\end{aligned}$$

Where y_{is} is the ppm for purchase i in state s

And mgstr_{is} , bulkp_{is} , and source_{is} are fixed effects

Linear mixed model fit by maximum likelihood ['lmerMod']

Formula: $\log(\text{ppm}) \sim \text{mgstr} + \text{bulk_purchase} + \text{source} + (1 \mid \text{state})$

Data: streetrx

AIC BIC logLik deviance df.resid

6366 6413 -3175 6350 2610

Scaled residuals:

Min 1Q Median 3Q Max

-9.143 -0.383 0.095 0.493 4.639

Random effects:

Groups Name Variance Std.Dev.

state (Intercept) 0.0124 0.111

Residual 0.6548 0.809

Number of obs: 2618, groups: state, 50

Fixed effects:

Estimate Std. Error t value

(Intercept) 0.1891 0.0501 3.77

mgstr10 -0.4345 0.0429 -10.12

mgstr40 -0.8240 0.0682 -12.08

bulk_purchase1 Bulk purchase -0.1325 0.0362 -3.67

sourceInternet -0.4013 0.0629 -6.38

sourcePersonal -0.1051 0.0343 -3.06

Correlation of Fixed Effects:

(Intr) mgst10 mgst40 bl_1Bp srcInt

mgstr10 -0.712

mgstr40 -0.465 0.519

blk_prch1Bp -0.195 0.010 0.035

sourcIntrnt -0.265 0.041 0.047 -0.010

sourcePrsnl -0.436 0.011 -0.002 -0.003 0.348

Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB
foo.c

```
clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include"
-DNDEBUG -I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/Rcpp/"
-I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/"
-I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/"
-I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/BH/include"
-I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/include/"
-I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/include/"
-I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppParallel/include/"
-I"/Library/Frameworks/R.framework/Versions/4.0/Resources/library/rstan/include"
-DEIGEN_NO_DEBUG -DBOOST_DISABLE_ASSERTS -DBOOST_PENDING_INTEGER
-DSTAN_THREADS -DBOOST_NO_AUTO_PTR -include "/Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/unsupported/Eigen/MatrixFunctions"
-D_REENTRANT -DRCPP_PARALLEL_USE_TBB=1 -I/usr/local/include
-fPIC -Wall -g -O2 -c foo.c -o foo.o
```

In file included from :1:

In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/include/StanHeaders/math/MatrixFunctions.hpp:

In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/MatrixFunctions:

In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/MatrixFunctions:

/Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/MatrixFunctions
error: unknown type name 'namespace'

```
namespace Eigen {
```

```
^
```

/Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/MatrixFunctions

error: expected ';' after top level declarator

```
namespace Eigen {
```

```
^
```

```
;
```

In file included from :1:

In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/include/StanHeaders/math/MatrixFunctions.hpp:

In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/MatrixFunctions:

/Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/MatrixFunctions

fatal error: 'complex' file not found

```
#include
```

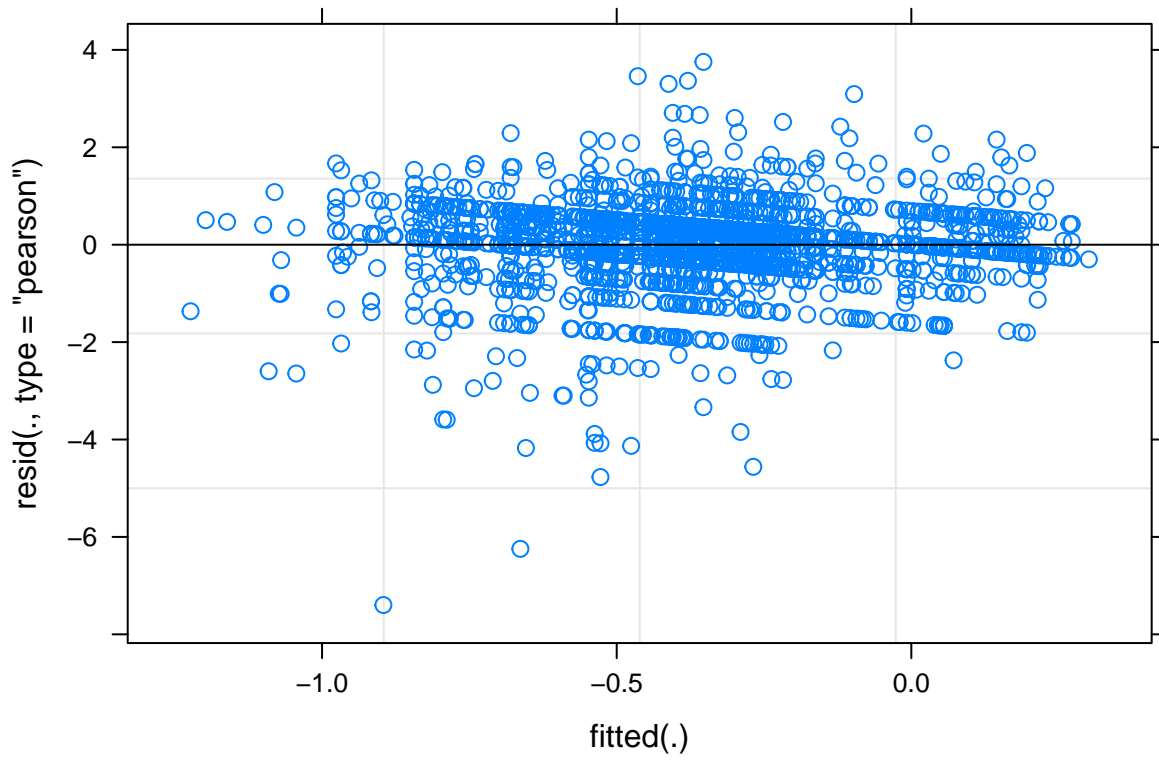
```
^~~~~~
```

3 errors generated.

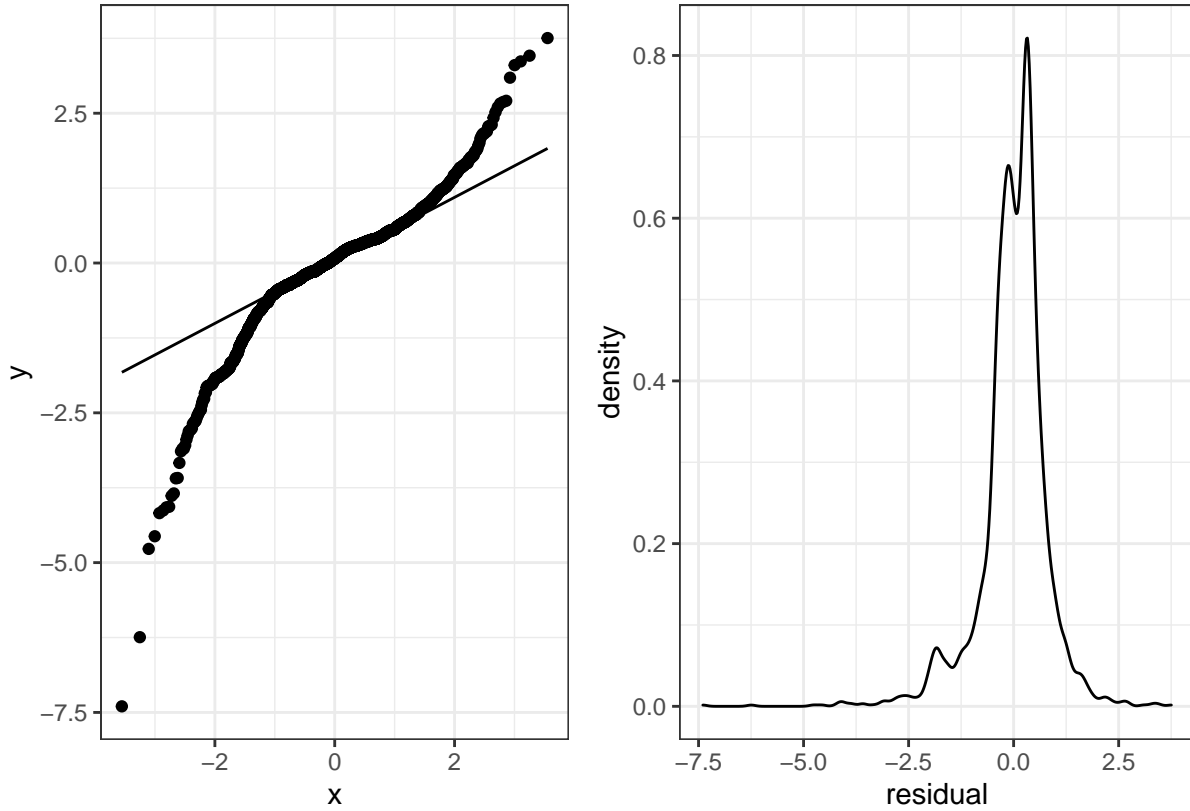
make: *** [foo.o] Error 1

“ One can see that the results given by Bayesian setting is almost the same is that from the frequentist setting.

Posterior checks:



Examine the normality of residuals



Interpretation of results

Result plots:

Fixed effects

Table 2: Fixed effect estimates

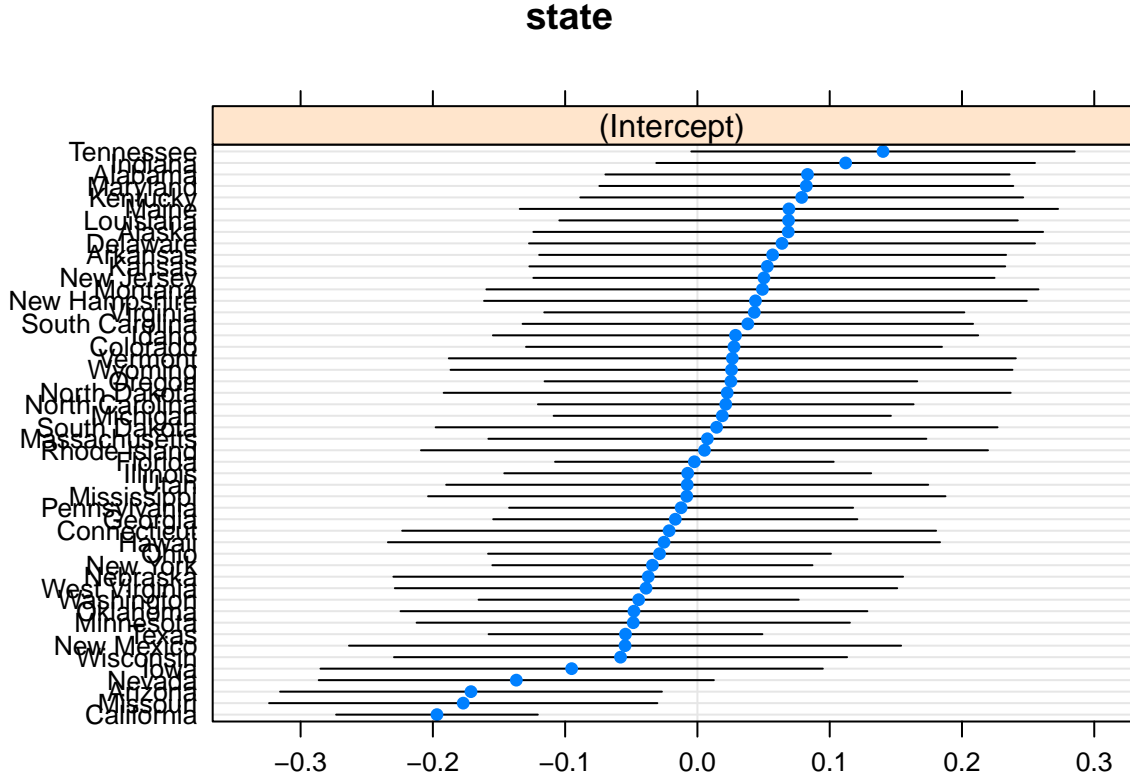
	Estimate	Std. Error	t value
(Intercept)	0.1891	0.0501	3.772
mgstr10	-0.4345	0.0429	-10.120
mgstr40	-0.8240	0.0682	-12.084
bulk_purchase1 Bulk purchase	-0.1325	0.0362	-3.665
sourceInternet	-0.4013	0.0629	-6.380
sourcePersonal	-0.1051	0.0343	-3.060

As we took the log of our response ppm, we must exponentiate our estimates in order to interpret the effect of each variable on ppm. The following results are:

- **Grand mean:** Our estimated grand mean for ppm is 1.21. This is the average price per mg of methadone across all states for dosages of 50mg, cases where the purchase was heard of, and non bulk purchases.
- **mgstr:** On average, we expect a decrease in dosage from 50 mg to 10 mg to result in a 35.24% decrease in price and decrease from 50 mg to 40 mg to yield a 56.13% decrease in price.
- **bulk_purchase:** On average bulk purchases are 12.41% cheaper, in terms of price per mg, than non bulk purchases.
- **source:** Purchases that were personally reported are on average 10% less price per mg than purchases that had been heard second hand, while we expect a purchase that was discovered through the internet to be 33% less price per mg than a purchase that had been heard second hand.

Random effects

The following illustrates the sorted estimated random state intercepts with 95% confidence intervals.



Overall the plot demonstrates significant heterogeneity across states in the baseline ppm of methadone for purchases that were in bulk and that had been reported by word of mouth. Three states, California, Missouri, and Arizona, do not contain 0 in their 95% confidence interval, implying a significant difference in their baseline ppm and the grand mean across all states. The plot also shows that the largest estimate was for Tennessee, 14.8% increase from the grand mean, while the smallest estimate was for California, a 18.13% decrease from the grand mean.

Table 3: Variance estimates

grp	var
state	0.0124
Residual	0.6548

Our across state variance of ppm is fairly small at 0.0124 while the within state variance remains large at 0.6548. Clearly, there is still a lot of within-group variance that our model is unable to explain.

Model Diagnostics (Influence of groups)

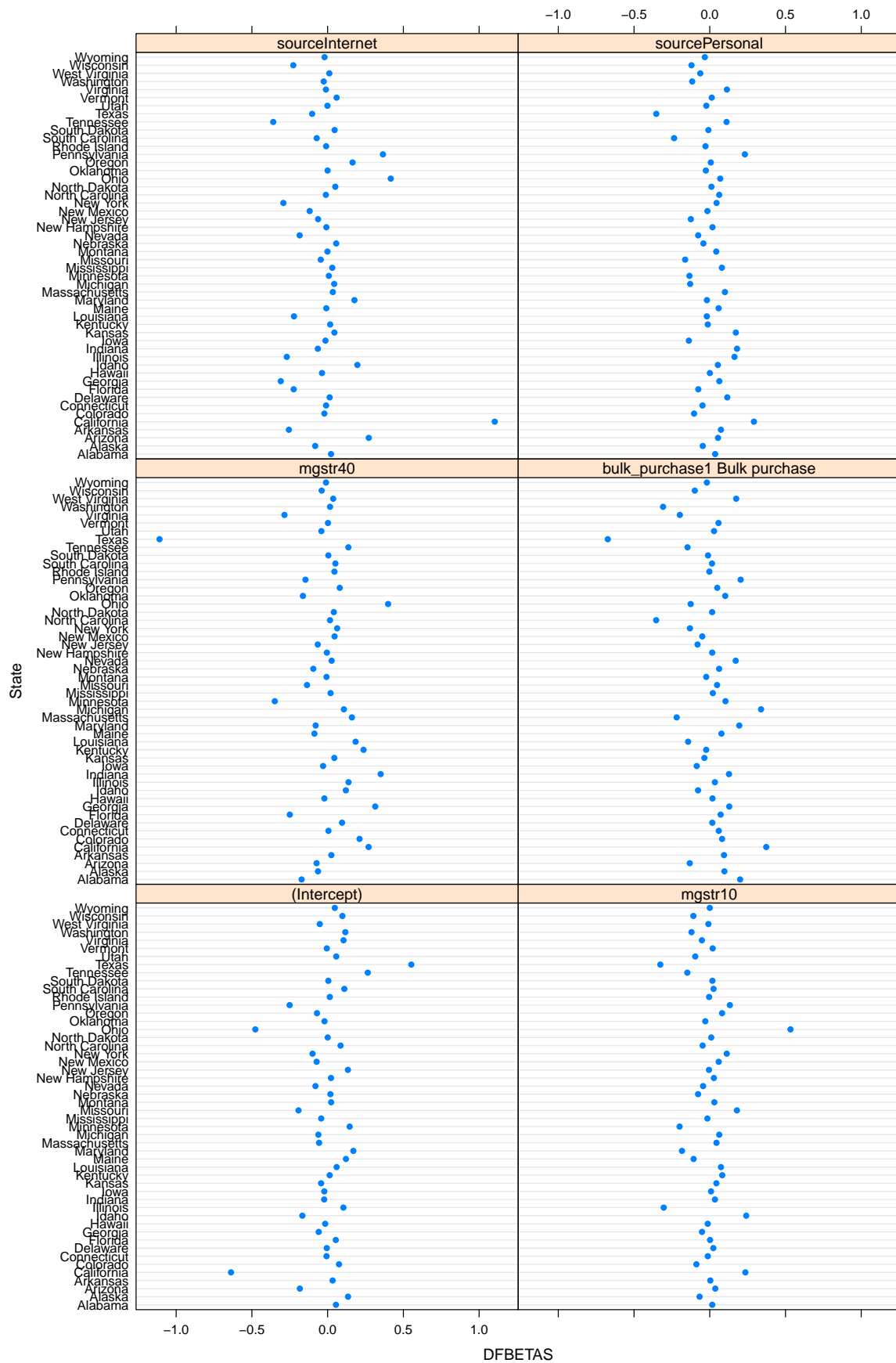
Cook's Distance

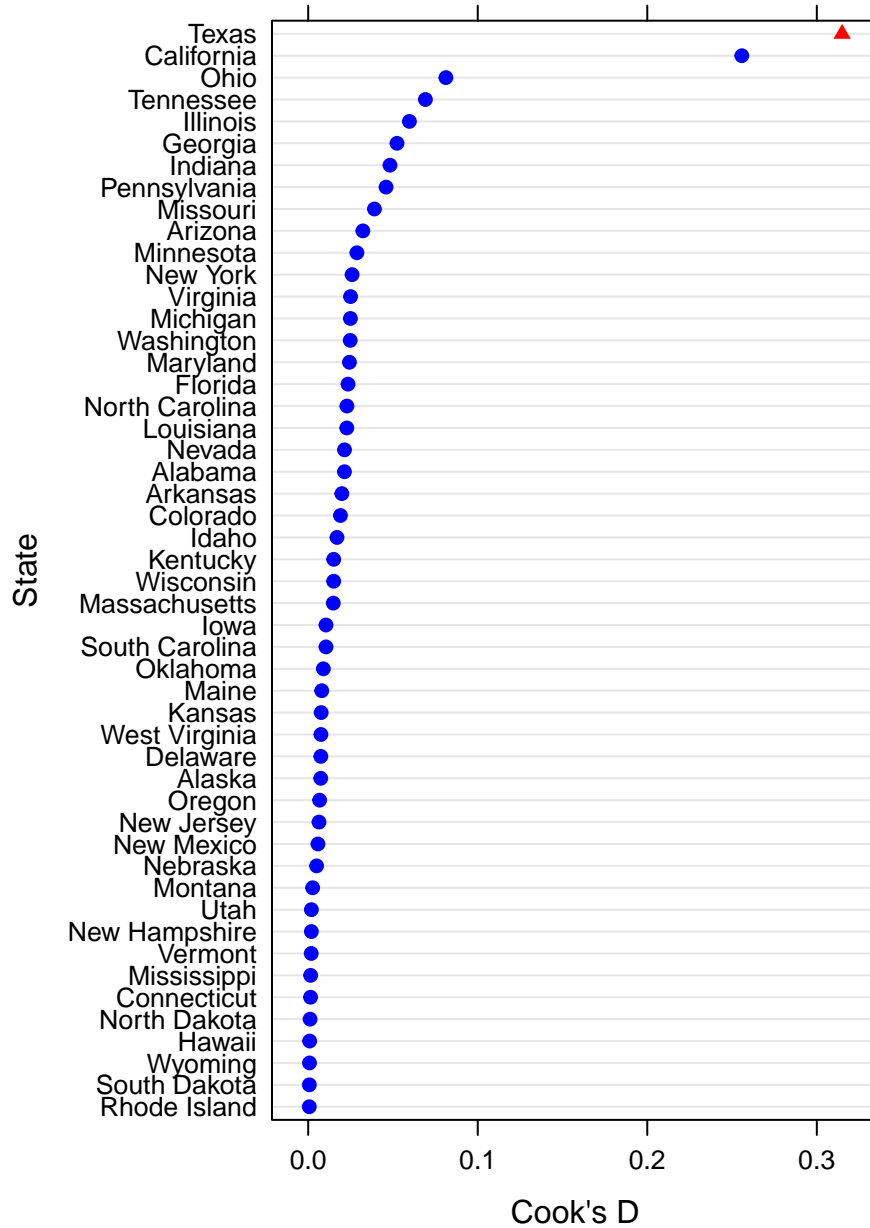
Now that we have our final model we will perform, model diagnostics to determine if there are influential groups that might be effecting our model assumptions. First we will look at the DFBETAS of each parameter for each state. Following this, we use Cook's distance as another criteria to determine if there are cases of influential states.

Table 4: Level of influences states have on single parameter estimates

	(Intercept)	mgstr10	mgstr40	bulk_purchase1 Bulk purchase	sourceInternet	sourcePersonal
California	-0.6379	0.2353	0.2705	0.3730	1.1032	0.2915
Georgia	-0.0590	-0.0511	0.3144	0.1285	-0.3095	0.0632
Illinois	0.1040	-0.3040	0.1379	0.0340	-0.2698	0.1632
Indiana	-0.0228	0.0341	0.3503	0.1270	-0.0648	0.1800
Michigan	-0.0611	0.0627	0.1073	0.3387	0.0433	-0.1294
Minnesota	0.1453	-0.1994	-0.3489	0.1038	0.0080	-0.1342
New York	-0.0996	0.1122	0.0628	-0.1307	-0.2923	0.0449
North Carolina	0.0851	-0.0472	0.0157	-0.3539	-0.0118	0.0624
Ohio	-0.4775	0.5329	0.3992	-0.1261	0.4166	0.0693
Pennsylvania	-0.2510	0.1329	-0.1463	0.2037	0.3651	0.2318
Tennessee	0.2651	-0.1482	0.1369	-0.1467	-0.3590	0.1108
Texas	0.5525	-0.3265	-1.1099	-0.6733	-0.1024	-0.3531
Virginia	0.1050	-0.0518	-0.2851	-0.1973	-0.0113	0.1132
Washington	0.1167	-0.1206	0.0159	-0.3083	-0.0258	-0.1160

The above table contains the states such that at least one parameter had a standardized difference in their estimate that exceeded our cutoff when excluding that state. Note that many of the states included have the largest sample sizes in the data set.





When examining Cook's distance, Texas is the only state that exceeds the cutoff and can be considered influential. This is not surprising as Texas had multiple parameters such that their DFBETAS exceeded the cutoff and Cook's distance is a summary measure of how an observation influences all parameter estimates. While, it seems Texas is an influence group we have yet to determine if it is an outlier. At the same time, Texas is has the third largest sample size in the data set and is an important data point in the analysis so it does not make sense to consider deleting this group.

Normality

Appendix

Additional EDA

Assessment of Random Slopes

Log(ppm) vs. mgstr, bulk purchase, and source by 8 random states

We do not observe difference in levels by state

