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 interatctions 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. Source contains links to websites where individuals purchased the Methadone, which will not be helpful in the model.

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.

Table 1: Variable Descriptions

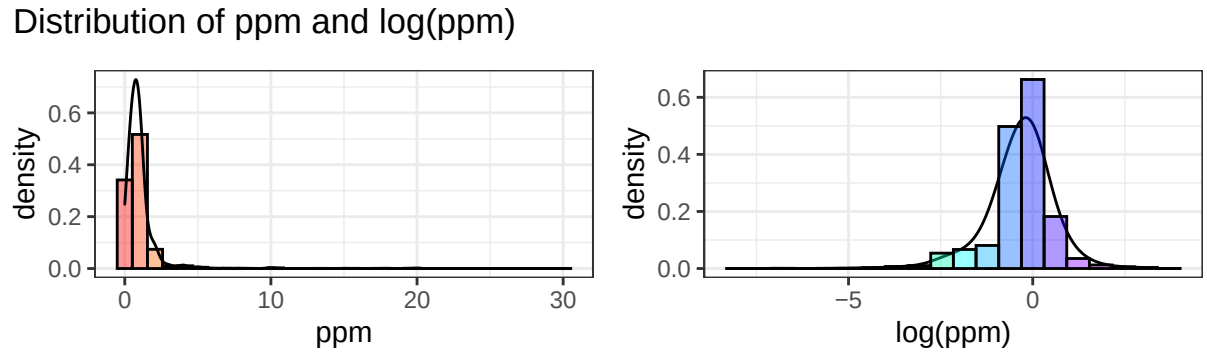
Variable	Description
ppm	Price per mg (outcome of interest)
yq_pdate	Year and quarter drug was purchased
price_date	Date of the reported purchase
city	city purchased
state	state purchased
country	country purchased
USA_region	northeast, midwest, west, south, or other/unknown
source	source of information
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)

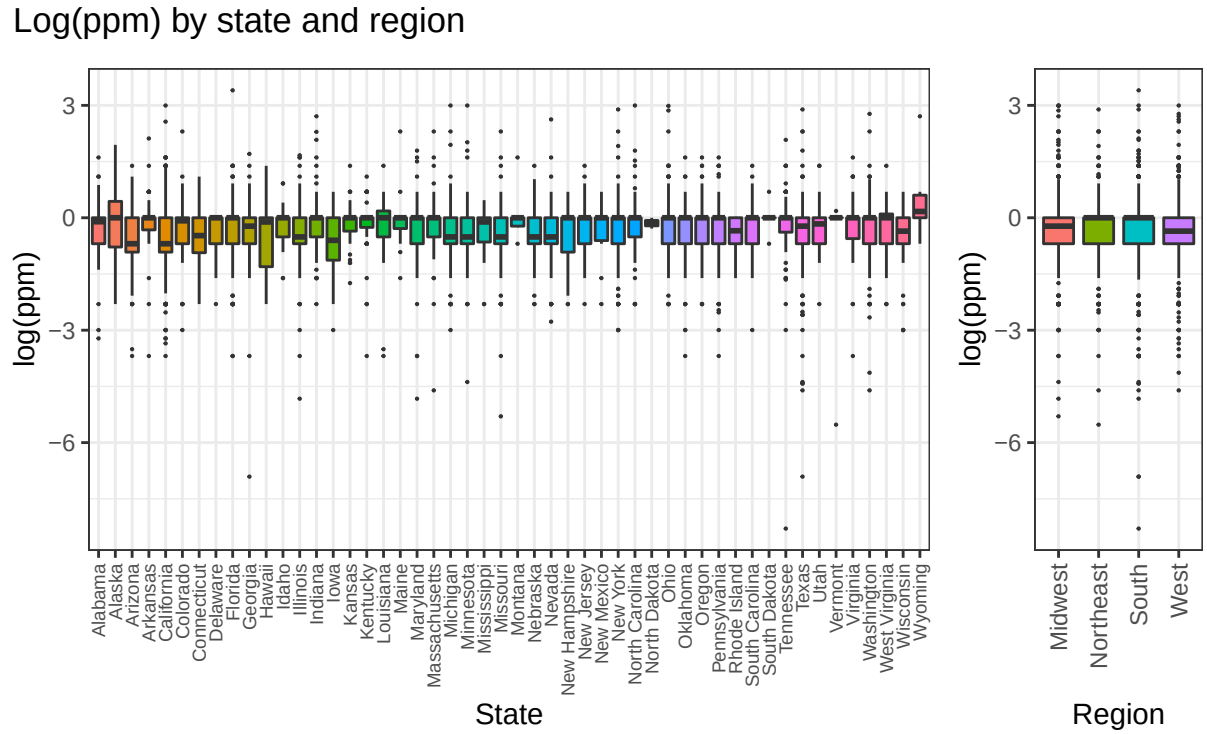
We explore the data first by examining the distribution of our outcome variable, price per mg (ppm). We next examine the distribution of our outcome variable, ppm (price per mg). We can see that ppm spans a large range of values, from \$0.00025/mg to \$40/mg. We note from the below histogram that the 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(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.



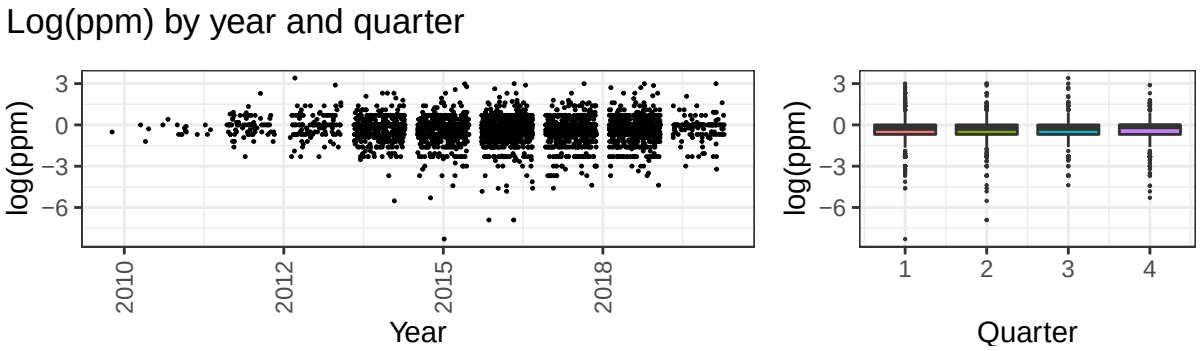
Assessing random intercepts

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(ppm) among both states and regions. From the below plots, we can see that there is not much variation of log(ppm) by region, but there is some variation by state. Thus, we will include a random intercept by state in our model.

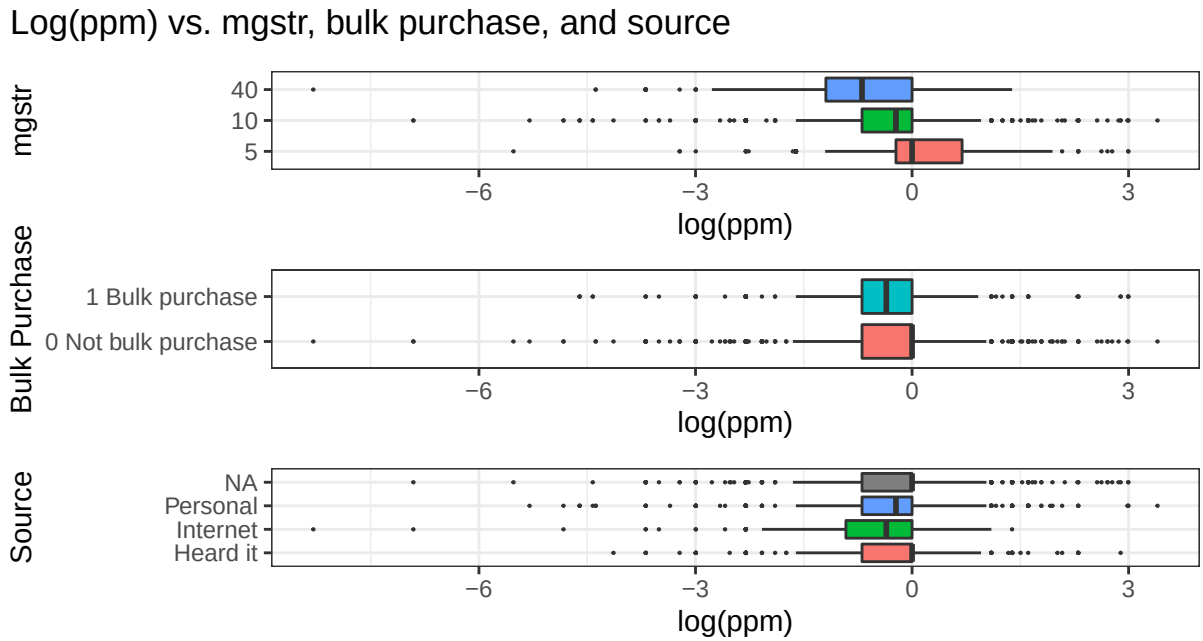


Assessing relationship of variables with log(ppm)

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.



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.

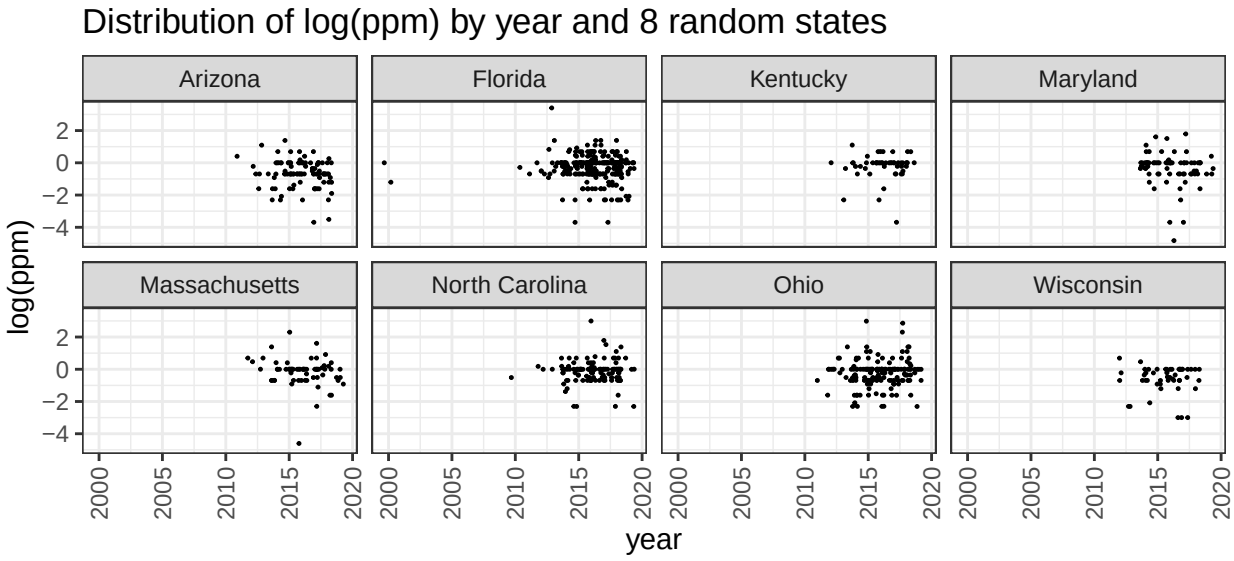


We note that form_temp (formulation of the drug as pill, patch, etc.) is always pill for Methadone, so we do not consider it as a potential variable in our model. The last variable in our dataset is primary reason for the purchase of Methadone. We note in the data cleaning section that this variable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



We then examine differences in the levels of mgstr, source, and bulk_purchase vs. log(ppm) by state. We created boxplots of each level of mgstr, source, and bulk_purchase vs. log(ppm) by state and did not find evidence of any major difference between the levels of the factors vs. log(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.

Model Building and Selection

I do not have much information about what the best model should be, given that we are still in early EDA. Nevertheless, a priori I have some preference on what to use / what not to use.

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 on a binary variable, we should put it as a grouping variable for ease to interpret.
- Other categorical variables are subject on how many levels we keep.

2) Base model

State: s.

$$y_{is} = \mu + \alpha_s + \epsilon_{is}.$$

Table 2: Exhaustive search of fixed effects using BIC

model	BIC
log(ppm) ~ mgstr + bulk_purchase + source + (1 | state)	6413
log(ppm) ~ mgstr + source + (1 | state)	6418
log(ppm) ~ year + mgstr + bulk_purchase + source + (1 | state)	6420
log(ppm) ~ year + mgstr + source + (1 | state)	6426
log(ppm) ~ bulk_purchase + source + (1 | state)	6558
log(ppm) ~ source + (1 | state)	6560
log(ppm) ~ year + bulk_purchase + source + (1 | state)	6562
log(ppm) ~ year + source + (1 | state)	6565
log(ppm) ~ mgstr + bulk_purchase + (1 | state)	10347
log(ppm) ~ mgstr + (1 | state)	10355
log(ppm) ~ year + mgstr + bulk_purchase + (1 | state)	10355
log(ppm) ~ year + mgstr + (1 | state)	10363
log(ppm) ~ bulk_purchase + (1 | state)	10582
log(ppm) ~ year + bulk_purchase + (1 | state)	10587
log(ppm) ~ 1 + (1 | state)	10587
log(ppm) ~ year + (1 | state)	10592

Table 3: Exhaustive search of fixed effects using BIC

model	BIC
log(ppm) ~ mgstr + bulk_purchase + source + (1 | state)	6413
log(ppm) ~ mgstr + bulk_purchase:source + (1 | state)	6418
log(ppm) ~ mgstr + source + bulk_purchase:source + (1 | state)	6418
log(ppm) ~ mgstr + bulk_purchase + source + bulk_purchase:source + (1 | state)	6418
log(ppm) ~ mgstr + bulk_purchase + bulk_purchase:source + (1 | state)	6418
log(ppm) ~ mgstr + source + (1 | state)	6418
log(ppm) ~ mgstr + year:bulk_purchase + year:source + (1 | state)	6420
log(ppm) ~ year + mgstr + year:bulk_purchase + year:source + (1 | state)	6420
log(ppm) ~ year + mgstr + source + year:bulk_purchase + (1 | state)	6420
log(ppm) ~ mgstr + source + year:bulk_purchase + (1 | state)	6420

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

See how much heterogeniety.

May compare the fitted variance of regional level means and state level means. If regional level means have larger variance, it might indicate a clustering effect within regions.

```
##      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.

Our best model is

$$y_{is} = \mu + \alpha_s + \beta_1 I(mgstr_{is} = 10) + \beta_2 I(mgstr_{is} = 40) + \beta_3 I(bulkp_{is} = 1) + \beta_4 I(source_{is} = internet) + \beta_4 I(source_{is} = personal) + \epsilon_{is}$$

$$\alpha_s \sim Normal(0, \tau^2)$$
$$\epsilon_{is} \sim Normal(0, \sigma^2)$$

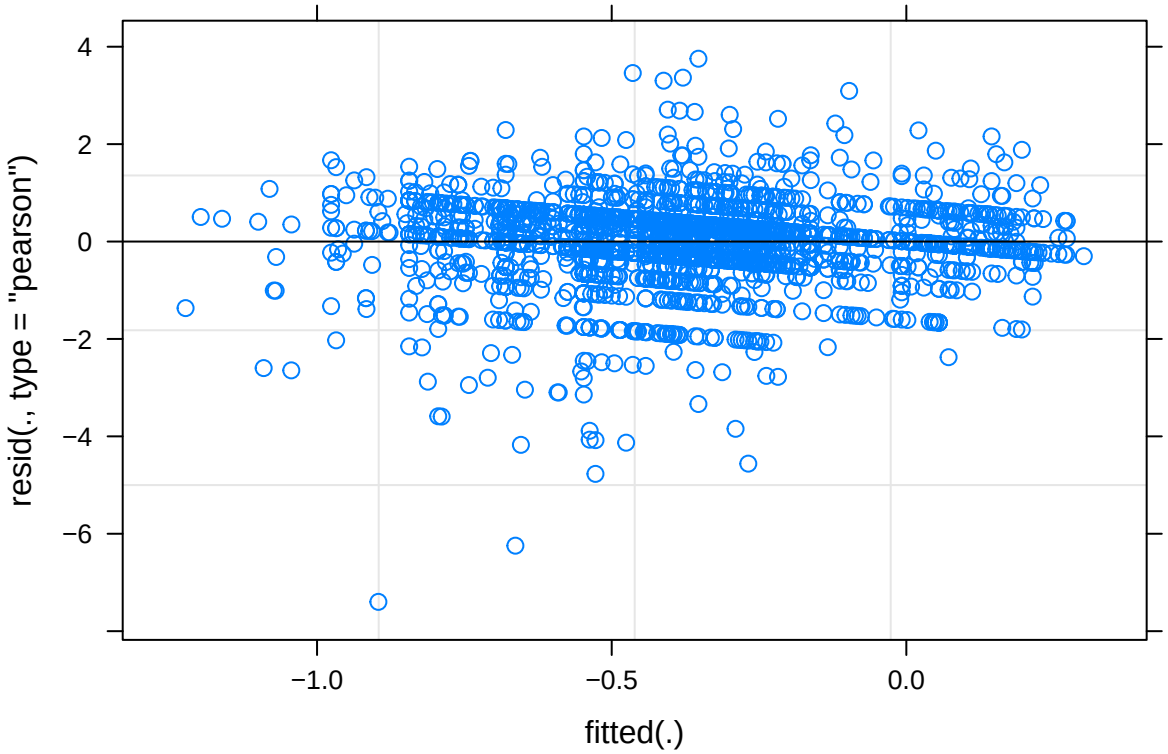
Where y_{is} is the ppm for purchase i in state s

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

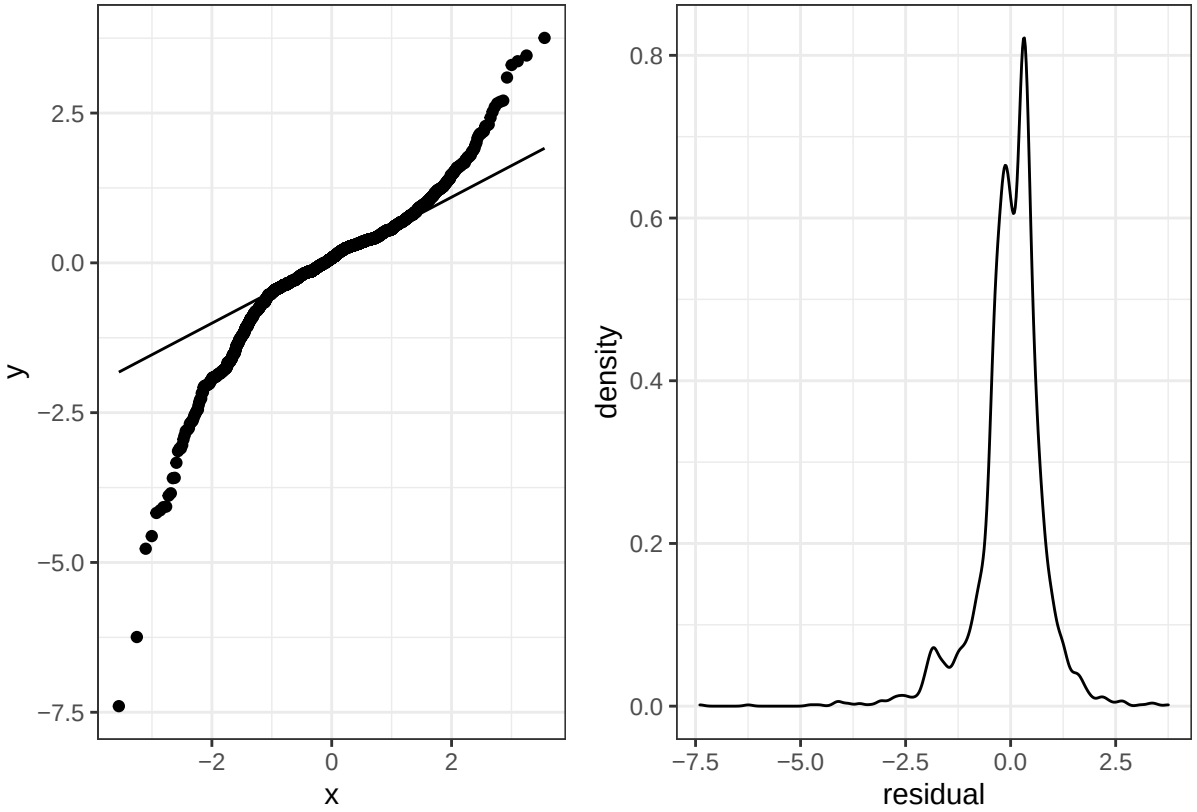
```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(ppm) ~ mgstr + bulk_purchase + source + (1 | state)
##    Data: streetrx
##
##      AIC      BIC   logLik deviance df.resid
##     6366     6413    -3175     6350     2610
##
## Scaled residuals:
##    Min     1Q Median     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
```

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 4: Fixed effect estimates

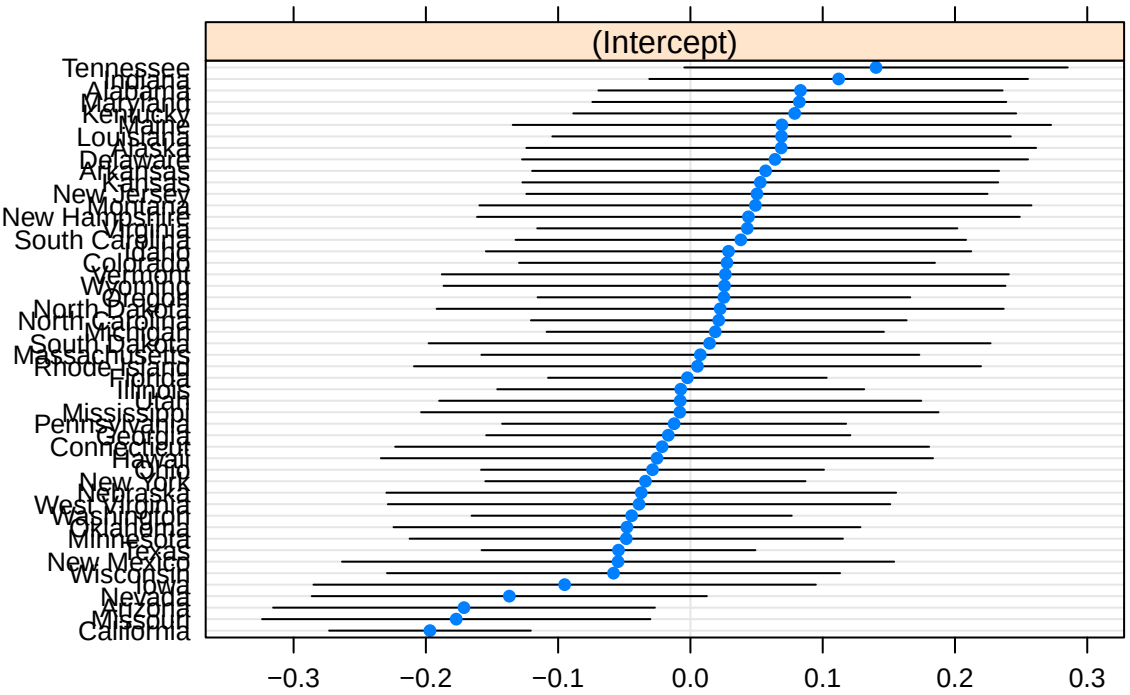
	Estimate	Std. Error	t value
(Intercept)	0.1891	0.0501	3.772
mgstr10	-0.4345	0.0429	-10.120
mgstr40	-0.8240	0.0682	-12.084
bulk_purchase1 Bulk purchase	-0.1325	0.0362	-3.665
sourceInternet	-0.4013	0.0629	-6.380
sourcePersonal	-0.1051	0.0343	-3.060

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_puchase`: 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 5: 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)

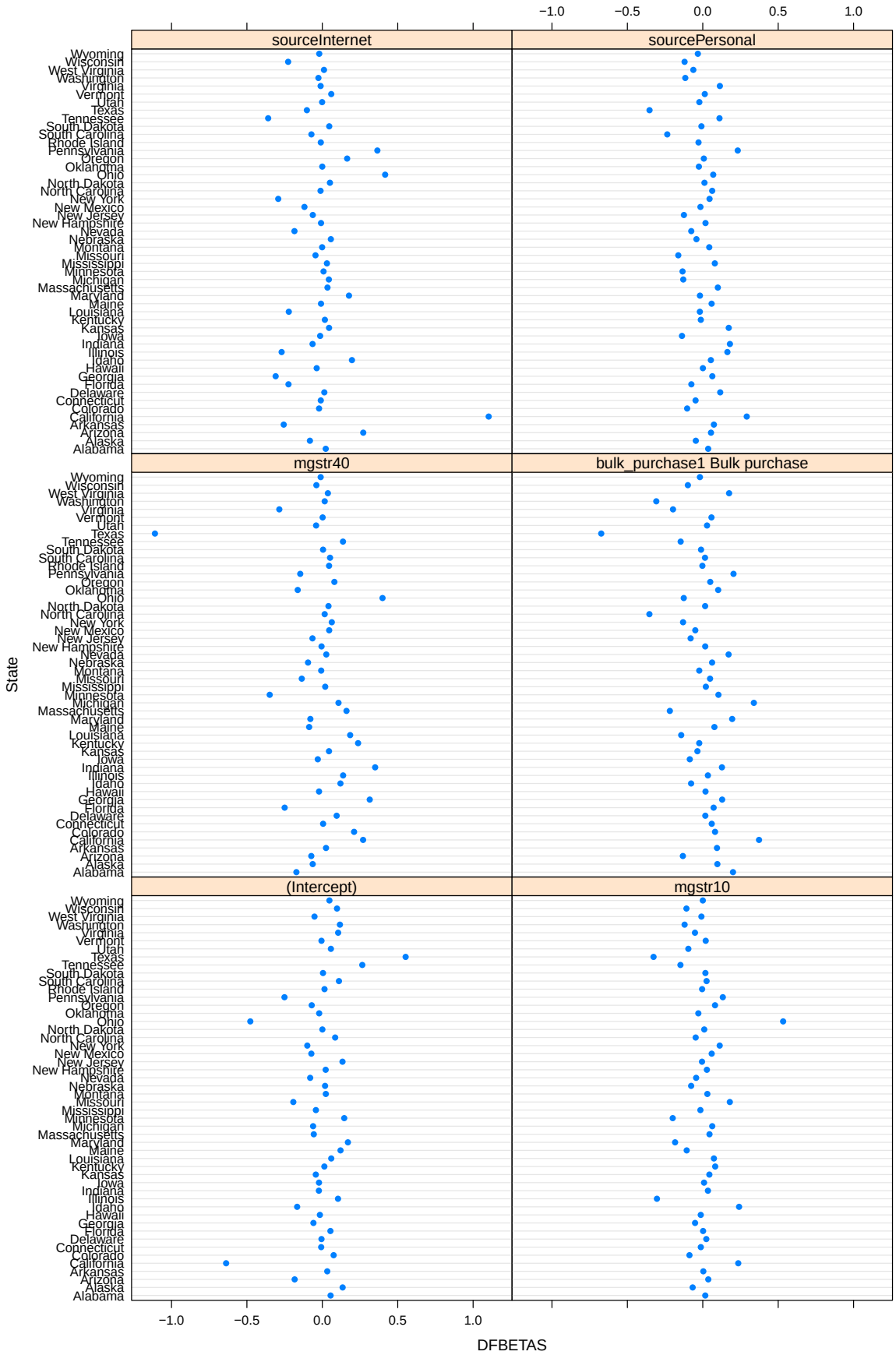
Cook's Distance

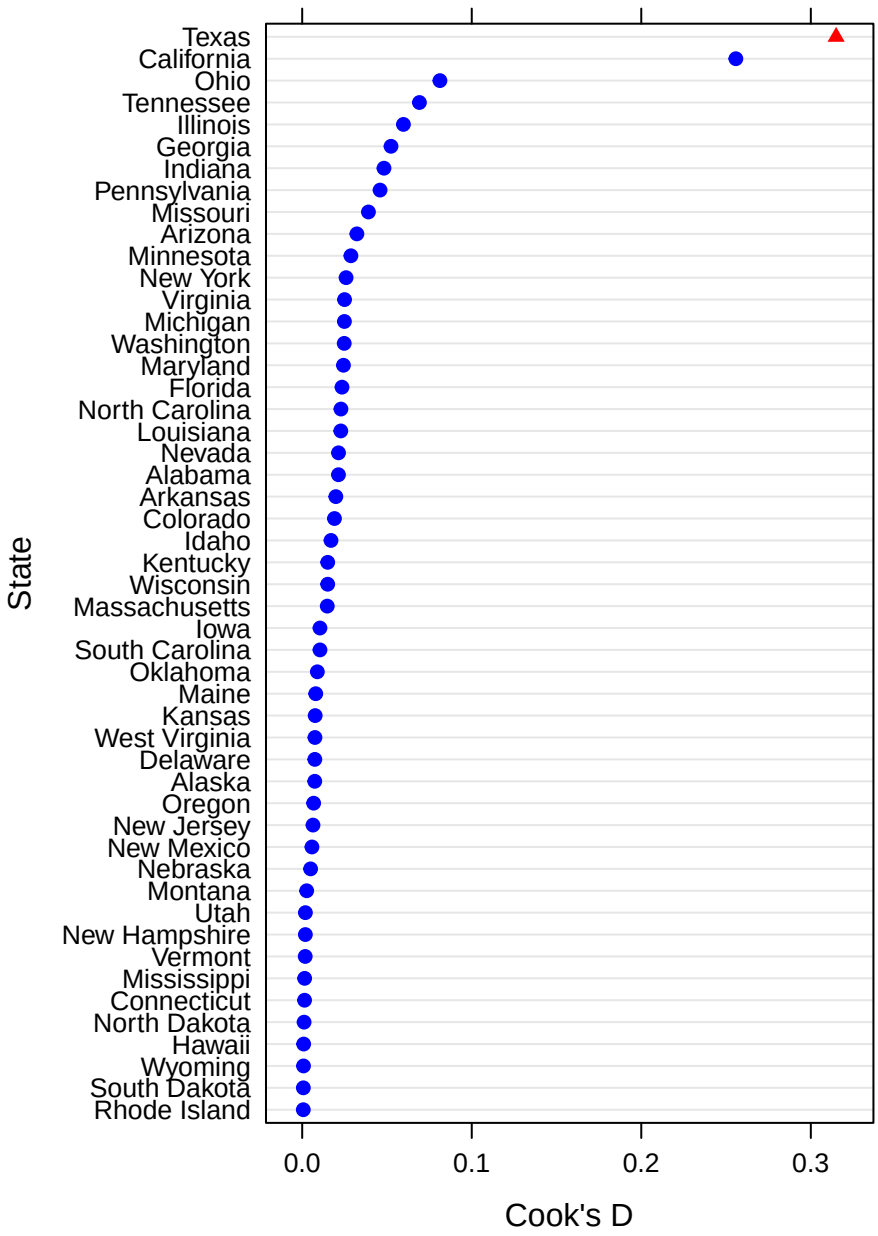
Now that we have our final model we will perform, model diagnostics to determine if there are influential groups that might be effecting our model assumptions. First we will look at the DFBETAS of each parameter for each state. Following this, we use Cook's distance as another criteria to determine if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

Table 6: Level of influences states have on single parameter estiamtes

	(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.


sourceInternet
sourcePersonal
mgstr40
bulk_purchase1 Bulk purchase
(Intercept)
mgstr10
State
DFBETAS
−1.0
−0.5
0.0
0.5
1.0
Wyoming
Wisconsin
West Virginia
Washington
Virginia
Vermont
Utah
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Dakota
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Nevada
Nebraska
Montana
Missouri
Mississippi
Minnesota
Michigan
Massachusetts
Maryland
Maine
Louisiana
Kentucky
Kansas
Iowa
Indiana
Illinois
Idaho
Hawaii
Georgia
Florida
Delaware
Connecticut
Colorado
California
Arkansas
Arizona
Alaska
Alabama



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