Case Study 1

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

10/17/2021

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 (in Table 1 below), and we will explore how the factors in the dataset are or are not associated with pricing per milligram.

Research Questions:

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

Data and Cleaning

Missing Data and Data Entry Errors We first examine missing data in the streetrx data. 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 over 50% of the observations are missing. 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.

Then, we proceed to modify variables and remove observation as needed. We note that all of the observations with a missing value for ppm are also missing in dosage (mgstr). We remove all data with a missing value for ppm (and thus mgstr as well), as it is nonsensical to include observations that are missing a value for the response variable. We thus have no missing mgstr values in the final data used for modeling, which is helpful.

Variable Modification From EDA we notice that most mgstr values are either 5, 10, or 40 mg, so we decide to remove the few observations that do not have these values (less than 10 observations). We then code mgstr as a factor variable as it has only three levels.

We also remove two observations in our data that have 1969 as year, which does not make sense given StreetRx did not exist at the time.

We remove observations that had input "USA" in the `state` variable, as this does not make sense, and would not provide information about heterogeneity in price per milligram values among states.

For the `source` variable, we combine all of the various websites that are reported and code them as "Internet". This greatly reduces the number of levels to just three.

We remove any states with only one observation, as it does not make sense to estimate within-group variance for states in this case. This removes only Washington D.C. and Guam, so we can still get useful insights into the heterogeneity present in ppm across all of the other states and territories in the dataset.

Finally, we also note that all purchases were made in the USA, so we do not consider the `country` variable.

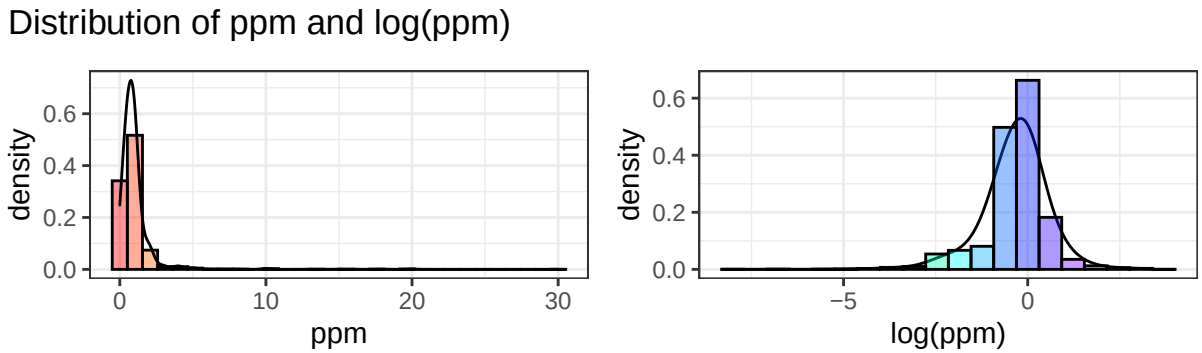
Table 1: Variable Descriptions

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

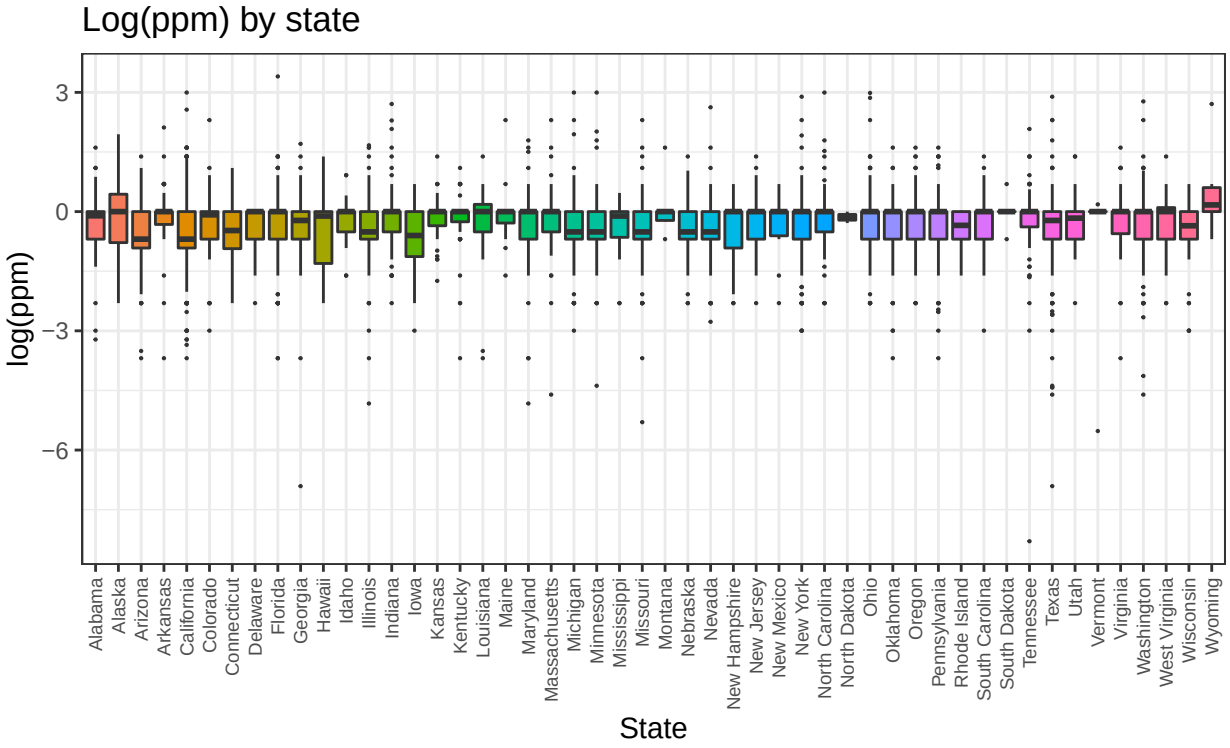
EDA

Distribution of outcome variable (ppm) We note from the below histogram that the distribution of ppm is highly right-skewed. To satisfy the conditional distribution assumption, we 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.

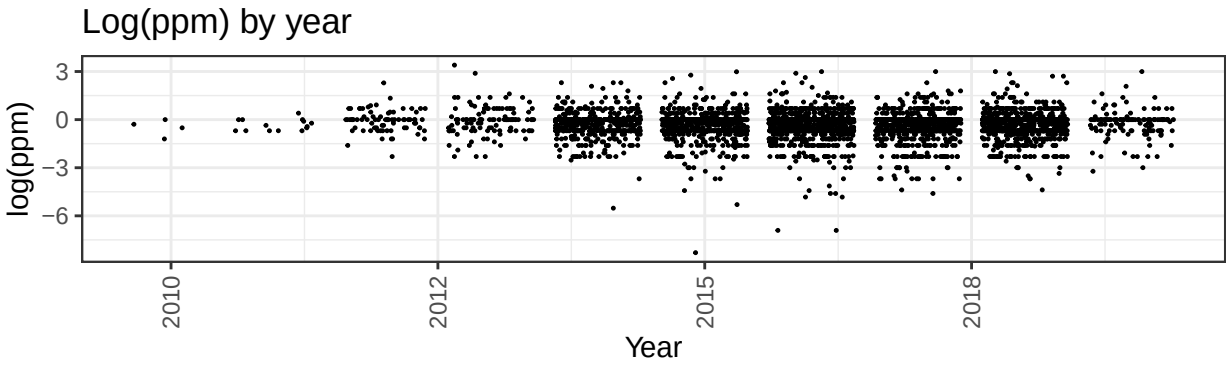
We choose to do a log transformation of ppm. We can that the histogram of log(ppm) is relatively normally distributed and symmetric. Using a log transformation is a good choice as well because our raw data ppm observation are all greater than 0, and log transformations are still easily interpretable,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



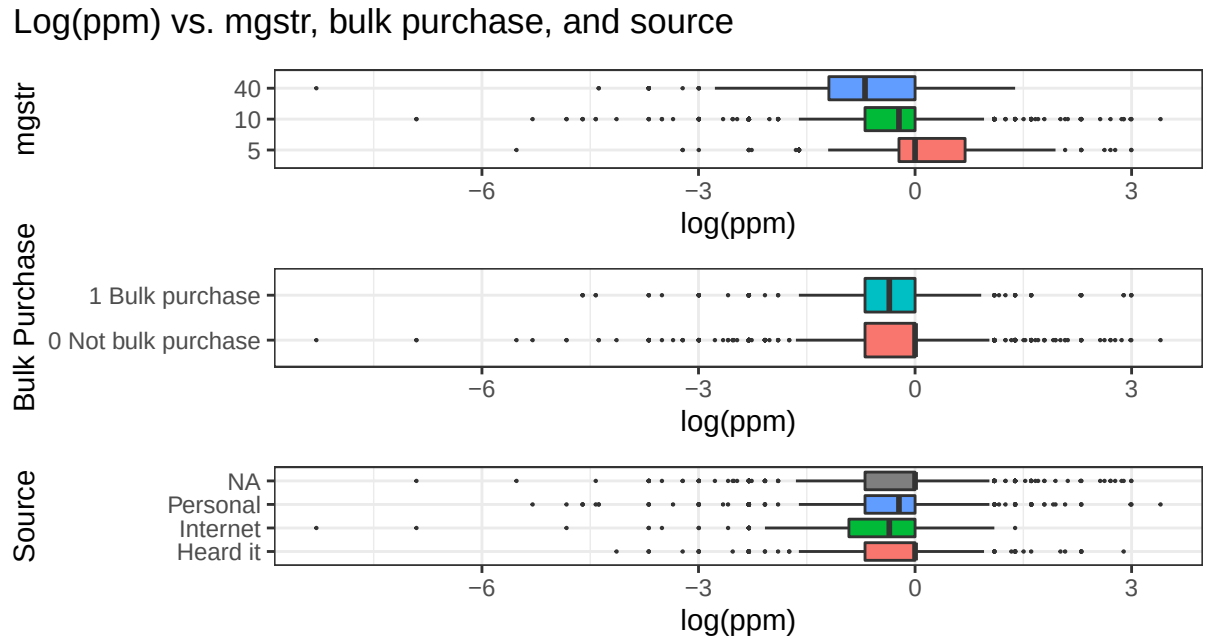
Assessing random intercepts One of our research questions is to understand heterogeneity in pricing of Methadone by location, which we will study by including a random intercept by a location variable. There are thus a few options for grouping variables for a random intercept. We exclude `city` as we 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 some variation by state, and not much for region (see boxplot in appendix). Thus, we will include a random intercept by state in our model.



Assessing relationship of variables with log(ppm) Our second research question implores us to understand which variables are associated with pricing per milligram of Methadone. We thus would like to include some fixed effects in our model that allow us to explore these associations. Through EDA, we assess relationships of variables in our dataset with log(ppm), which 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 that quarter does not seem to have variation by log(ppm), so we do not include this in our model (see boxplot in appendix). 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. To evaluate random slopes through EDA, 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, and find that there is no distinguishable difference so we choose not to include a random slope of year by state (see plot in appendix).

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.

Model Building and Selection

Model Selection 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`. To simplify the modeling while still being able to interpret uncertainty, we decide to use a two phase process. First, we run an exhaustive search on all models that include our desired fixed/main effects, random intercept for `state`, and all 2-way interactions. This gives $2^10 = 1024$ models to select from, and we then select the model in the lowest BIC as our final model. We do note that many models that are fit through exhaustive search do not have good interpretability (ex. models that include a two-way interaction but exclude the corresponding main effects). Thus, we fit all 1024 models, but select the final model with the smallest BIC among all the models that include all main effects where a two-way interaction exists. The table below shows the top 10 models with the lowest BIC from our exhaustive search.

Table 2: 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

Our best model with the lowest BIC is:

$$y_{is} = \mu + \alpha_s + \beta_1 \mathbb{I}(\text{mgstr}_{is} = 10) + \beta_2 \mathbb{I}(\text{mgstr}_{is} = 40) + \beta_3 \mathbb{I}(\text{bulkp}_{is} = 1) + \beta_4 \mathbb{I}(\text{source}_{is} = \text{internet}) + \beta_4 \mathbb{I}(\text{source}_{is} = \text{personal}) + \epsilon_{is},$$

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

where y_{is} is the ppm for purchase i in state s, and mgstr_{is}, bulkp_{is}, and source_{is} are fixed effects.

Frequentist vs. Bayesian Model In the 2nd phase of our model build, we run both a frequentist and Bayesian hierarchical model to estimate the parameters, including the coefficients and the variances. We found that estimates from the Bayesian version of our model were very similar to the frequentist model we fitted, so we proceeded to interpret the frequntist model that we fitted above. The Bayesian model estimates and posterior distributions can be found in the appendix.

Bayesian Model Comparison

Model Fitting

Here we impose a prior on the parameters and fit the model using a Bayesian approach to see if we have differing estimates from the frequentist version. Since we do not have much information about the model, we used non-informative priors. We choose not to employ an empirical bayes approach here, as we want to be careful about using the data to inform our priors.

$$\beta_j \overset{iid}{\sim} \text{Normal}(0, 1)$$
$$\tau^2 \sim \text{InvGamma}(0.1, 0.1)$$
$$\sigma^2 \sim \text{InvGamma}(0.1, 0.1)$$

Estimates Comparison In the below table, we compare the point estimates of the parameters from the frequentist model and posterior means of the parameters from the Bayesian model. We can see that the results given in the Bayesian setting are almost the same as those from the frequentist setting. We show the plots of the posterior distributions for all parameters in the appendix, and note that the traceplots suggest that the sampling chains converge.

Table 3: Fixed effect estimates comparison

	Frequentist Model Estimate	Bayesian Model Esimate
(Intercept)	0.1891	0.1870
mgstr10	-0.4345	-0.4325
mgstr40	-0.8240	-0.8213
bulk_purchase1 Bulk purchase	-0.1325	-0.1313
sourceInternet	-0.4013	-0.3985
sourcePersonal	-0.1051	-0.1041

Table 4: Variance estimates comparison

Source of Variation	Bayesian Model Estimate	Frequentist Model Estimate
Residual	0.6568	0.6548
State	0.0139	0.0124

Interpretation of results (frequentist model)

Fixed effects

Table 5: Fixed effect estimates on log scale

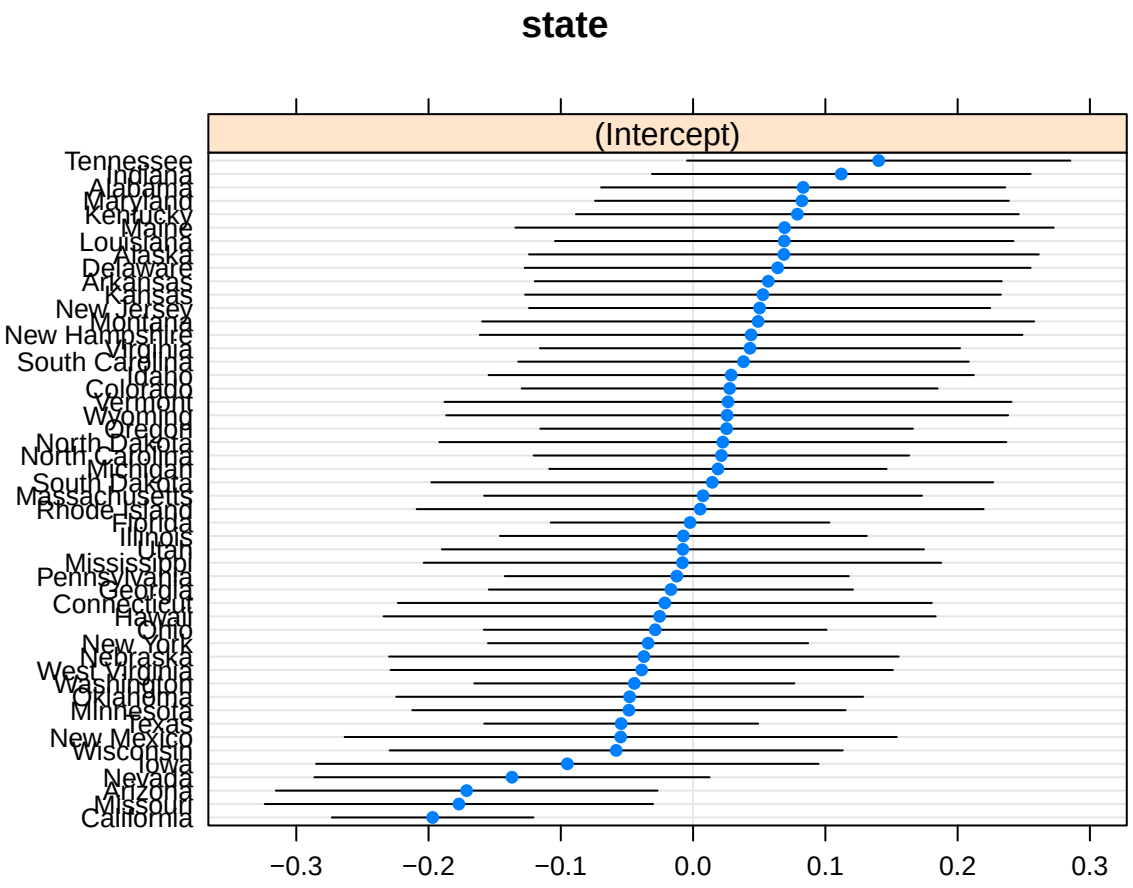
	Estimate	Std. Error	t value
(Intercept)	0.1891	0.0501	3.772
mgstr10	-0.4345	0.0429	-10.120
mgstr40	-0.8240	0.0682	-12.084
bulk_purchase1 Bulk purchase	-0.1325	0.0362	-3.665
sourceInternet	-0.4013	0.0629	-6.380
sourcePersonal	-0.1051	0.0343	-3.060

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 6: 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 and Limitations

There are a few limitations to note with our model. As mentioned above, our across-state variance is much smaller than within-state variance, meaning that there is still much within-group variance that our model is unable to explain. Additionally, there are issues with the normality assumption as well as outliers explained below.

Residual Analysis Through examining our residuals we can determine how well our model assumptions hold. The plot of our residual against the fitted values are somewhat reassuring (see appendix). Aside for a couple points that appear to be potential outliers, we observed that the constant variance and linearity conditions are met. It is possible that the two outliers contribute to our larger within state variance, though this is something we do not explore in this analysis.

The normality assumption is not met, as the distribution of the residuals appears to have fatter tails, particularly on the left side of the distribution where a few outliers are present. With that being said, the density of the residuals is fairly symmetric and centered at 0 (see appendix for plot).

Influential Groups Now we take steps to determine if there are influential groups that might be effecting our model assumptions. First we 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.

There are 14 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. These include California, Texas, and New York. A table of states and their corresponding DFBETAS for each variable can be seen in the appendix.

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 with DFBETAS that 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 influential group we have yet to determine if it is an outlier as well. At the same time, Texas has the third largest sample size in the data set and clearly a state we would want to include in the analysis so it does not make sense to consider deleting this group.

Appendices

We include both a Code Appendix and a Plot Appendix

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE,
                      fig.align = 'center')
library(tidyverse)
library(lme4)
library(rstan)
library(brms)
library(knitr)
library(kableExtra)
library(patchwork)
library(lubridate)
library(gridExtra)
library(influence.ME)

options(scipen = 0, digits = 4)
ggplot2::theme_set(ggplot2::theme_bw())
#load data
load("streetrx.RData")

# GROUP 1 - METHADONE
# filter data for only methadone
streetrx <- streetrx %>%
  filter(api_temp == "methadone")
# variable description table
tibble(Variable = names(streetrx),
Description = c("Price per mg (outcome of interest)",
                "   Year and quarter drug was purchased",
                "Date of the reported purchase
",
                "city purchased",
                "state purchased",
"country purchased",
"northeast, midwest, west, south, or other/unknown
",
"source of information",
"active ingredient of drug of interest, in our case Methadone)
",
"formulation of the drug (e.g., pill, patch, suppository)
",
"dosage strength in mg of the units purchased
",
"indicator for purchase of 10+ units at once
",
"primary reason for purchase"
)) %>%
  kable(caption = "Variable Descriptions") %>%
  kable_styling(latex_options = "HOLD_position")
# code missing data
```

```r
streetrx <- data.frame(apply(streetrx, 2, function(x) gsub("^$|^ $", NA, x)))

# table of variables with missing data
# tibble(
#   Variable = names(streetrx),
#   `Number Missing` = apply(streetrx, 2, function(x) sum(is.na(x))),
#   `Proportion Missing` = apply(streetrx, 2, function(x) mean(is.na(x)))
# ) %>%
#   kable(caption = "Number of observations missing per variable") %>%
#   kable_styling(latex_options = "HOLD_position")

# code factors and numeric variables
streetrx <- streetrx %>%
  mutate(ppm = as.numeric(ppm),
         yq_pdate = as.numeric(yq_pdate),
         price_date = mdy(price_date),
         city = as.factor(city),
         state = as.factor(state),
         country = as.factor(country),
         USA_region = as.factor(USA_region),
         source = as.factor(source),
         form_temp = as.factor(form_temp),
         mgstr = as.numeric(mgstr),
         bulk_purchase = as.factor(bulk_purchase),
         Primary_Reason = as.factor(Primary_Reason)
         )

# add year
streetrx <- streetrx %>%
  mutate(year = year(price_date))

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

# delete levels of mgstr that are not 5, 10, 40
streetrx <- streetrx %>%
  filter(mgstr %in% c(5,10,40)) %>%
  mutate(mgstr = as.factor(mgstr))

# delete year of 1969 as it is likely an error
streetrx <- streetrx %>%
  filter(year != 1969)

# delete state = USA
streetrx <- streetrx %>%
  filter(state != "USA")

#combine all website sources as being Internet source
streetrx <- streetrx %>%
  mutate(source = as.character(source)) %>%
  mutate(source = if_else(str_detect(source, "http://"), "Internet", source)) %>%
  mutate(source = if_else(str_detect(source, ".com$"), "Internet", source)) %>%
```

```
  mutate(source = if_else(source == "Streetrx", "Internet",source)) %>%
  mutate(source = if_else(source ==  "Poopy,", "N/A", source)) %>%
  mutate(source = if_else(source == "google", "Internet", source)) %>%
  mutate(source = if_else(source == "Internet Pharmacy", "Internet", source)) %>%
  mutate(source = na_if(source, "N/A")) %>%
  mutate(source = na_if(source, "None")) %>%
  mutate(source = as.factor(source))

# remove states with 1 observation
streetrx <- streetrx %>%
  group_by(state) %>%
  count() %>%
  filter(n > 1) %>%
  select(state) %>%
  left_join(streetrx, by = "state")
ppm_hist1 <- streetrx %>%
  ggplot(aes(x = ppm, y = ..density..)) +
  geom_histogram(alpha = 0.4, fill=rainbow(30), bins=30, color = "black") +
  geom_density(color = "black", adjust = 5)

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

patchwork <- ppm_hist1 + ppm_hist2

patchwork + plot_annotation(
  title = "Distribution of ppm and log(ppm)")
ggplot(streetrx, aes(x = state, y = log(ppm))) +
  geom_boxplot(aes(fill = factor(state)), outlier.size = 0.1) +
  labs(title = "Log(ppm) by state",
       x = "State") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size = 7))
streetrx %>%
  mutate(
    year = yq_pdate %/% 10,
    quarter = yq_pdate %% 10
  ) %>%
  filter(year > 2009) %>%
  ggplot(., aes(x = year, y = log(ppm))) +
  geom_jitter(size = 0.2) +
  labs(title = "Log(ppm) by year",
       x = "Year") +
  theme(legend.position = "none",
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
mgstr_ppm <- ggplot(streetrx, aes(x = factor(mgstr), y = log(ppm))) +
    geom_boxplot(aes(fill = factor(mgstr)), outlier.size = 0.1) +
  labs(x = "mgstr") +
  theme(legend.position = "none") +
  coord_flip()
```

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

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

patchwork <- mgstr_ppm / bp_ppm / source_ppm

patchwork + plot_annotation(
  title = "Log(ppm) vs. mgstr, bulk purchase, and source")
exhaustive_search <- function(raw_model, vars, data, REML = F) {
  y_name <- deparse(raw_model[[2]])
  group_name <- deparse(raw_model[[3]])
  id <- 0 : (2^length(vars) - 1)
  construct_model <- function(.id){
    subset <- (.id %/% 2^(0:(length(vars) - 1))) %% 2 == 1
    if (all(subset == F)){
      RHS <- paste0(c("1", group_name), collapse = ' + ')
    }
    else {
      RHS <- paste0(c(vars[subset], group_name), collapse = ' + ')
    }
    paste(y_name, RHS, sep = ' ~ ')
  }
  run_model <- function(.id){
    model_str <- construct_model(.id)
    model_formula <- as.formula(model_str)
    res <- lmer(model_formula, REML = REML, data = data)
    return (summary(res)$AICtab)
  }

  bind_cols(
    model = sapply(id, construct_model),
    as.data.frame(t(sapply(id, run_model)))
  )
}
ex_result_int <- exhaustive_search(
  raw_model <- log(ppm) ~ (1 | state),
  vars = c("year", "mgstr", "bulk_purchase", "source",
           "year:mgstr", "year:bulk_purchase", "year:source",
           "mgstr:bulk_purchase", "mgstr:source", "bulk_purchase:source"),
  data = streetrx,
  REML = F
)
ex_result_int %>% arrange(BIC) %>% head(10) %>%
  select(1, 3) %>%
  kbl(caption = "Exhaustive search of fixed effects using BIC") %>%
```

```
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = "HOLD_position")
best_model <- ex_result_int %>% arrange(BIC) %>% .[1, "model"]
res <- lmer(as.formula(best_model), REML = F, data = streetrx)
summary(res)
priors <- c(
  set_prior("normal(0, 1)", class = 'b'),
  set_prior("inv_gamma(0.1, 0.1)", class = "sd", group = "state"),
  set_prior("inv_gamma(0.1, 0.1)", class = 'sd')
)

bayes_result <- brm(as.formula(best_model), data = streetrx, prior = priors,
                    verbose = F, refresh = 0)

bayes_fixed <- (summary(bayes_result)$fixed) %>%
  select("Estimate") %>%
  rename("Bayesian Model Esimate" = "Estimate")

freq_est <- as.matrix(coef(summary(res))[,1])
colnames(freq_est)[[1]] <- "Frequentist Model Estimate"



cbind(freq_est, bayes_fixed) %>%
  knitr::kable(caption = "Fixed effect estimates comparison") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = "HOLD_position")

bayesian_var <- rbind(summary(bayes_result)$spec_pars %>%
  select("Estimate"),
  summary(bayes_result)$random$state %>% select("Estimate")) %>%
  mutate(Estimate = Estimate^2) %>%
  rename("Bayesian Model Estimate" = "Estimate")

rownames(bayesian_var) <- c("Residual", "State")

frequentist_est <- as.data.frame(VarCorr(res)) %>%
  select(1, 4) %>%
  rename(var = vcov) %>%
  rename("Frequentist Model Estimate" = var) %>%
  select("Frequentist Model Estimate")


rownames(frequentist_est) <- c("State", "Residual")

merge(bayesian_var, frequentist_est, by = "row.names") %>%
  rename("Source of Variation" = "Row.names") %>%
  knitr::kable(caption = "Variance estimates comparison") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = "HOLD_position")
coef(summary(res)) %>%
  knitr::kable(caption = "Fixed effect estimates on log scale") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
```

```
                latex_options = "HOLD_position")
# dotplot
library(lattice)
dotplot(ranef(res, condVar = TRUE), font.size = 5, rotate = TRUE)$state
as.data.frame(VarCorr(res)) %>%
  select(1, 4) %>%
  rename(var = vcov) %>%
  knitr::kable(caption = "Variance estimates") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = "HOLD_position")

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

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

streetrx %>%
  filter(yq_pdate >= 20000) %>%
  mutate(days = difftime(price_date, '2000-01-01', units = "days")) %>%
  ggplot() +
  geom_point(aes(x = days, y = log(ppm)))

streetrx %>%
  filter(yq_pdate >= 20090) %>%
  mutate(days = difftime(price_date, '2000-01-01', units = "days")) %>%
  ggplot() +
  geom_point(aes(x = days, y = log(ppm)))
###### GEOGRAPHICAL INFORMATION ######

ggplot(streetrx, aes(x = USA_region, y = log(ppm))) +
  geom_boxplot(aes(fill = factor(USA_region))) +
  labs(title = "Relationship between region and log(ppm)",
       x = "Region") +
  theme(legend.position = "none")

###### form_temp ######
streetrx %>% group_by(form_temp) %>% summarize(n = n())
streetrx %>% group_by(form_temp, is.na(mgstr)) %>% summarize(n = n())
# All syrup/liquid rows do not have `mgstr` values. Basically we do not have any outcome data for syrup
###### Quarter
streetrx %>% mutate(quarter = yq_pdate %% 10) %>%
  ggplot(aes(fill = factor(quarter))) +
  geom_boxplot(aes(x = quarter, y = log(ppm), group = quarter), outlier.size = 0.1) +
  theme(legend.position = "none") +
  labs(x = "Quarter")
```

```r
set.seed(7)
state_rand <- streetrx %>%
  group_by(state) %>%
  count() %>%
  arrange(desc(n)) %>%
  filter(n > 30) %>%
  ungroup() %>%
  sample_n(8) %>%
  pull(state)

state_rand <- as.character(state_rand)

streetrx %>%
  filter(state %in% c(state_rand)) %>%
  ggplot(., aes(x = year, y = log(ppm))) +
  geom_jitter(size = 0.2) +
  facet_wrap(~state, ncol = 4) +
  labs(title = "Distribution of log(ppm) by year and 8 random states") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
set.seed(7)
state_rand <- streetrx %>%
  group_by(state) %>%
  count() %>%
  arrange(desc(n)) %>%
  filter(n > 30) %>%
  ungroup() %>%
  sample_n(8) %>%
  pull(state)

state_rand <- as.character(state_rand)

mgstr_rand_slope <- streetrx %>%
  filter(state %in% c(state_rand)) %>%
  ggplot(., aes(x = mgstr, y = log(ppm), fill = mgstr)) +
  geom_boxplot() +
  facet_wrap(~state, ncol = 4) +
  theme(legend.position = "none") +
  coord_flip()
set.seed(0)
state_rand <- streetrx %>%
  group_by(state) %>%
  count() %>%
  arrange(desc(n)) %>%
  filter(n > 30) %>%
  ungroup() %>%
  sample_n(8) %>%
  pull(state)

state_rand <- as.character(state_rand)

source_rand_slope <- streetrx %>%
  filter(state %in% c(state_rand)) %>%
  ggplot(., aes(x = source, y = log(ppm), fill = source)) +
```

```
  geom_boxplot() +
  facet_wrap(~state, ncol = 4) +
  theme(legend.position = "none") +
  coord_flip()
set.seed(99)
state_rand <- streetrx %>%
  group_by(state) %>%
  count() %>%
  arrange(desc(n)) %>%
  filter(n > 30) %>%
  ungroup() %>%
  sample_n(8) %>%
  pull(state)

state_rand <- as.character(state_rand)

bp_rand_slope <- streetrx %>%
  filter(state %in% c(state_rand)) %>%
  ggplot(., aes(x = bulk_purchase, y = log(ppm), fill = bulk_purchase)) +
  geom_boxplot() +
  facet_wrap(~state) +
  theme(legend.position = "none") +
  coord_flip()
patchwork <- mgstr_rand_slope / source_rand_slope / bp_rand_slope

patchwork + plot_annotation(
  title = "Log(ppm) vs. mgstr, bulk purchase, and source by 8 random states",
  subtitle = 'We do not observe difference in levels by state',
)
ggplot(streetrx) +
  geom_boxplot(aes(x = bulk_purchase, y = log(ppm))) +
  facet_wrap(~mgstr)
ggplot(streetrx) +
  geom_boxplot(aes(x = source, y = log(ppm))) +
  facet_wrap(~mgstr)
ggplot(streetrx) +
  geom_boxplot(aes(x = bulk_purchase, y = log(ppm))) +
  facet_wrap(~source)
streetrx %>%
  mutate(quarter = yq_pdate %% 10) %>%
  ggplot(aes(x = as.factor(quarter), y = log(ppm))) +
  geom_boxplot() +
  facet_wrap(~ bulk_purchase)
streetrx %>%
  mutate(quarter = yq_pdate %% 10) %>%
  ggplot(aes(x = as.factor(quarter), y = log(ppm))) +
  geom_boxplot() +
  facet_wrap(~ mgstr)
streetrx %>%
  mutate(quarter = yq_pdate %% 10) %>%
  ggplot(aes(x = as.factor(quarter), y = log(ppm))) +
  geom_boxplot() +
  facet_wrap(~ source)
```

```
streetrx %>%
  mutate(year = yq_pdate %/% 10) %>%
  ggplot(aes(x=year, y = log(ppm)))  +
  geom_point() +
  facet_wrap(~mgstr) +
  geom_smooth(method = "lm")

streetrx %>%
  mutate(year = yq_pdate %/% 10) %>%
  filter(year > 2005) %>%
  ggplot(aes(x=year, y = log(ppm)))  +
  geom_point() +
  facet_wrap(~mgstr) +
  geom_smooth(method = "lm")
streetrx %>%
  mutate(year = yq_pdate %/% 10) %>%
  ggplot(aes(x=year, y = log(ppm)))  +
  geom_point() +
  facet_wrap(~bulk_purchase) +
  geom_smooth(method = "lm")


streetrx %>%
  mutate(year = yq_pdate %/% 10) %>%
  filter(year > 2005) %>%
  ggplot(aes(x=year, y = log(ppm)))  +
  geom_point() +
  facet_wrap(~bulk_purchase) +
  geom_smooth(method = "lm")
streetrx %>%
  mutate(year = yq_pdate %/% 10) %>%
  ggplot(aes(x=year, y = log(ppm)))  +
  geom_point() +
  facet_wrap(~source) +
  geom_smooth(method = "lm")

streetrx %>%
  mutate(year = yq_pdate %/% 10) %>%
  filter(year > 2005) %>%
  ggplot(aes(x=year, y = log(ppm)))  +
  geom_point() +
  facet_wrap(~source) +
  geom_smooth(method = "lm")
best_model <- ex_result_int %>% arrange(BIC) %>% .[1, "model"]
res <- lmer(as.formula(best_model), REML = F, data = streetrx)
summary(res)
priors <- c(
  set_prior("normal(0, 1)", class = 'b'),
  set_prior("inv_gamma(0.1, 0.1)", class = "sd", group = "state"),
  set_prior("inv_gamma(0.1, 0.1)", class = 'sd')
)

bayes_result <- brm(as.formula(best_model), data = streetrx, prior = priors,
```

```
                    verbose = F, refresh = 0)

bayes_fixed <- (summary(bayes_result)$fixed) %>%
  select("Estimate") %>%
  rename("Bayesian Model Esimate" = "Estimate")

freq_est <- as.matrix(coef(summary(res))[,1])
colnames(freq_est)[[1]] <- "Frequentist Model Estimate"



cbind(freq_est, bayes_fixed) %>%
  knitr::kable(caption = "Fixed effect estimates comparison") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = "HOLD_position")

bayesian_var <- rbind(summary(bayes_result)$spec_pars %>%
  select("Estimate"),
  summary(bayes_result)$random$state %>% select("Estimate")) %>%
  mutate(Estimate = Estimate^2) %>%
  rename("Bayesian Model Estimate" = "Estimate")

rownames(bayesian_var) <- c("Residual", "State")

frequentist_est <- as.data.frame(VarCorr(res)) %>%
  select(1, 4) %>%
  rename(var = vcov) %>%
  rename("Frequentist Model Estimate" = var) %>%
  select("Frequentist Model Estimate")


rownames(frequentist_est) <- c("State", "Residual")

merge(bayesian_var, frequentist_est, by = "row.names") %>%
  rename("Source of Variation" = "Row.names") %>%
  knitr::kable(caption = "Variance estimates comparison") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                latex_options = "HOLD_position")
plot(bayes_result)
plot(res)
residual <- resid(res)
p1 <- ggplot() +
  geom_qq(aes(sample = residual)) +
  geom_qq_line(aes(sample = residual)) +
  coord_equal()
p2 <- ggplot() +
  geom_density(aes(x = residual))
p1 + p2
best_model.inf <- influence(res, "state")
cutoff <- 2/sqrt(length(unique(streetrx$state)))

dfbetas_inf <- round(dfbetas(best_model.inf), 4)
above_cutoff <- apply(abs(dfbetas_inf) > cutoff, MARGIN = 1, any)
```

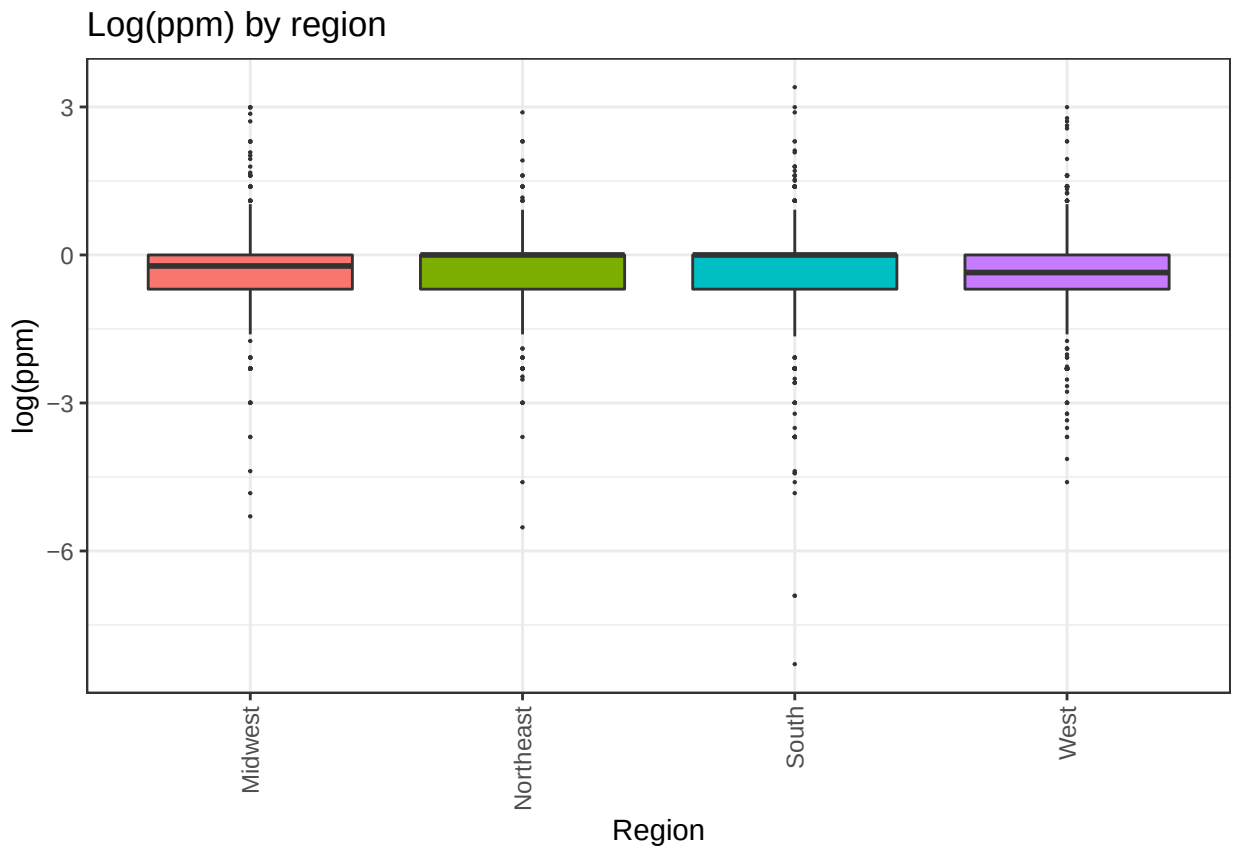
```
dfbetas_inf[above_cutoff,] %>%
  knitr::kable(caption = "Level of influence states have on single parameter estiamtes") %>%
  kable_styling(latex_options = "HOLD_position")

plot(best_model.inf,which="dfbetas",xlab="DFBETAS",ylab="State")
plot(best_model.inf,which="cook",cutoff=cutoff,
sort=TRUE,xlab="Cook's D",ylab="State")
```

Plot Appendix

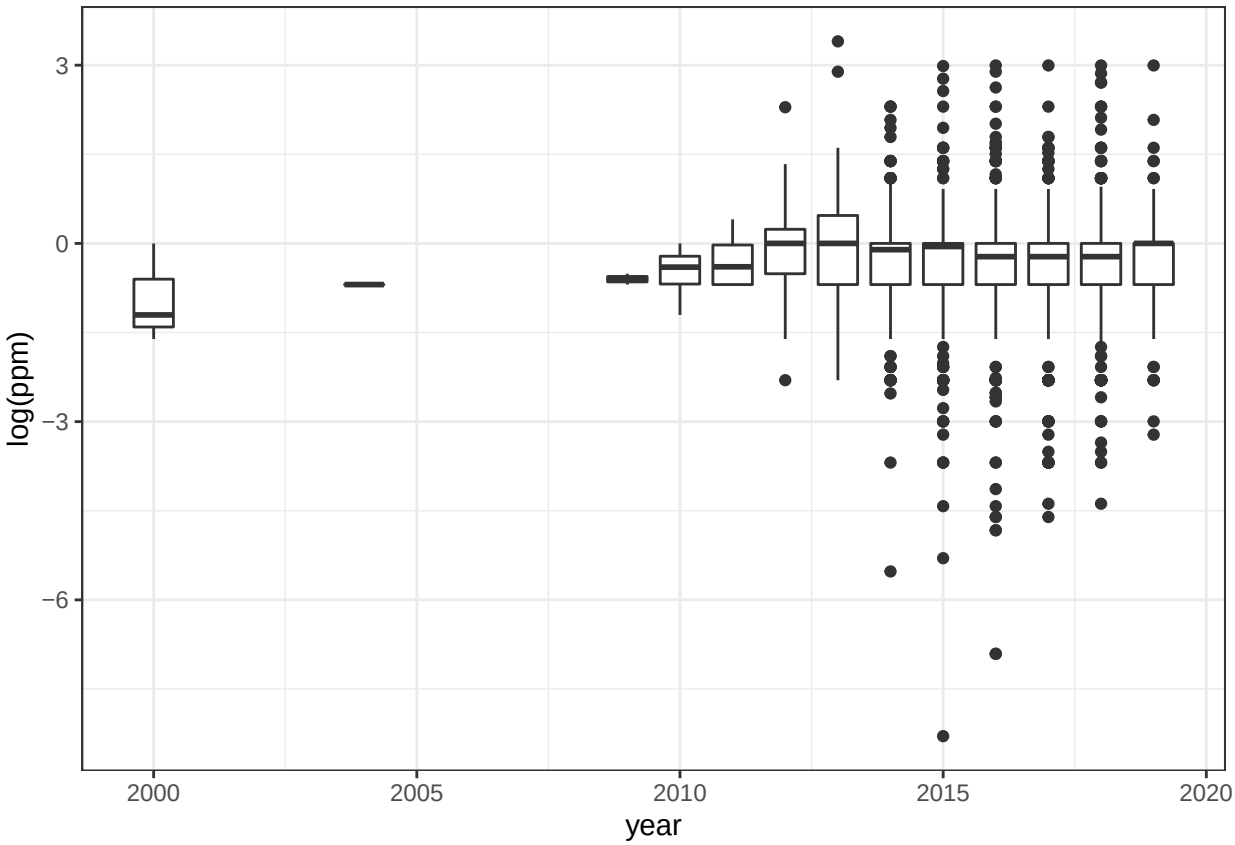
Additional EDA

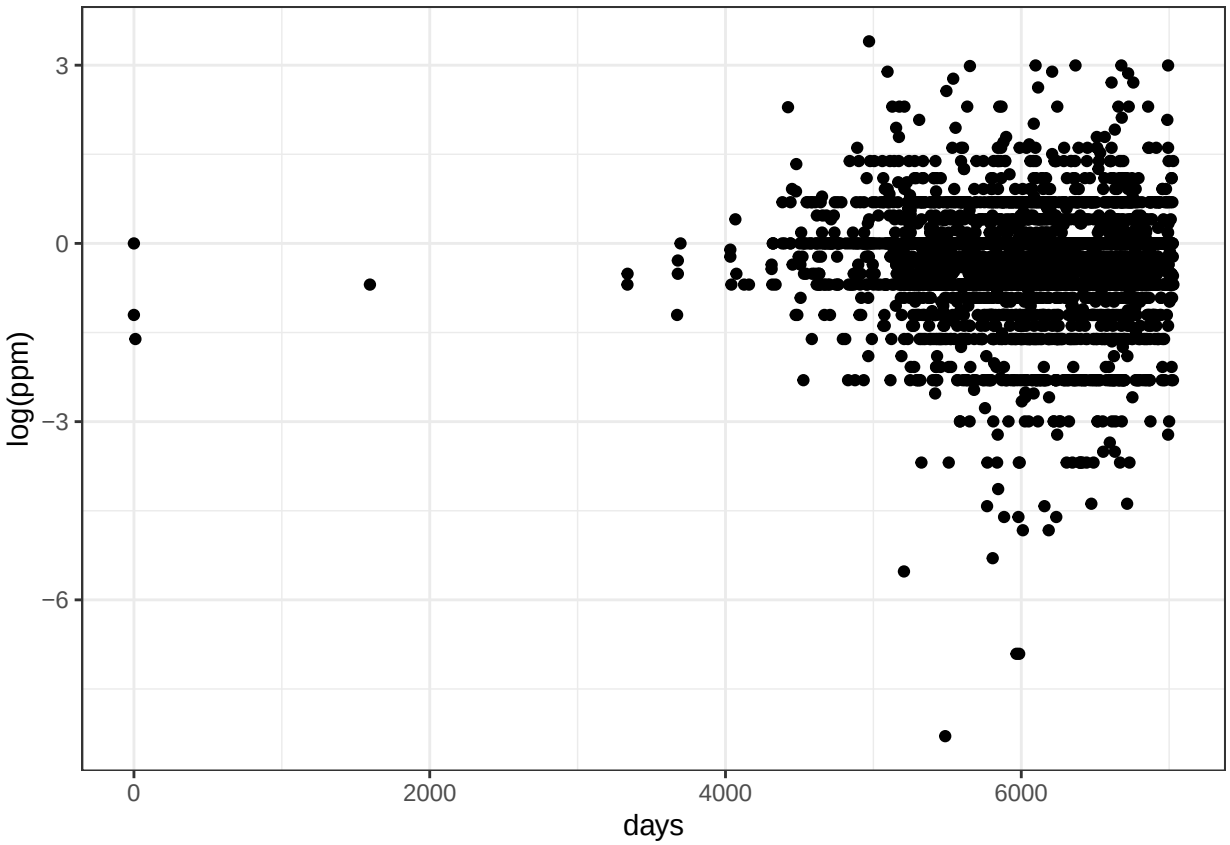
Random Intercepts

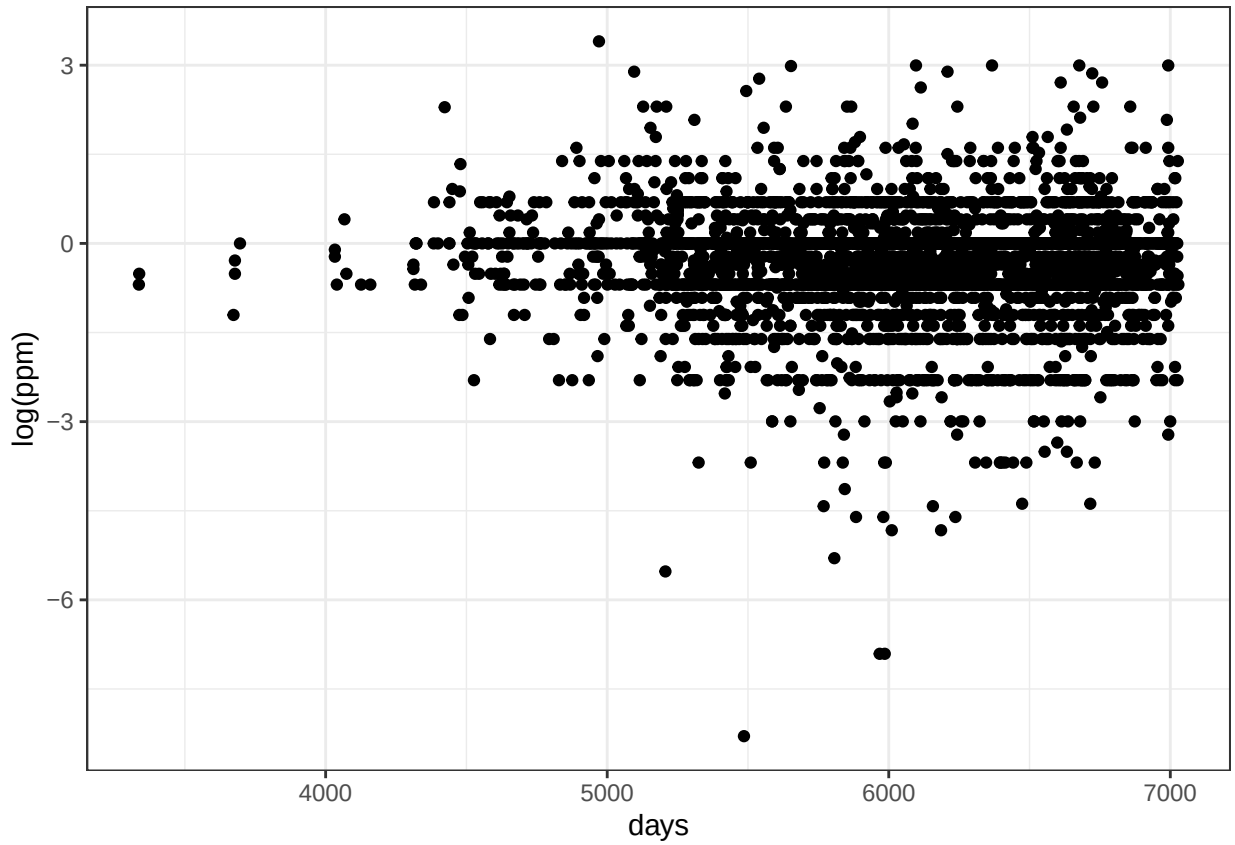


Fixed effects vs. log(ppm)

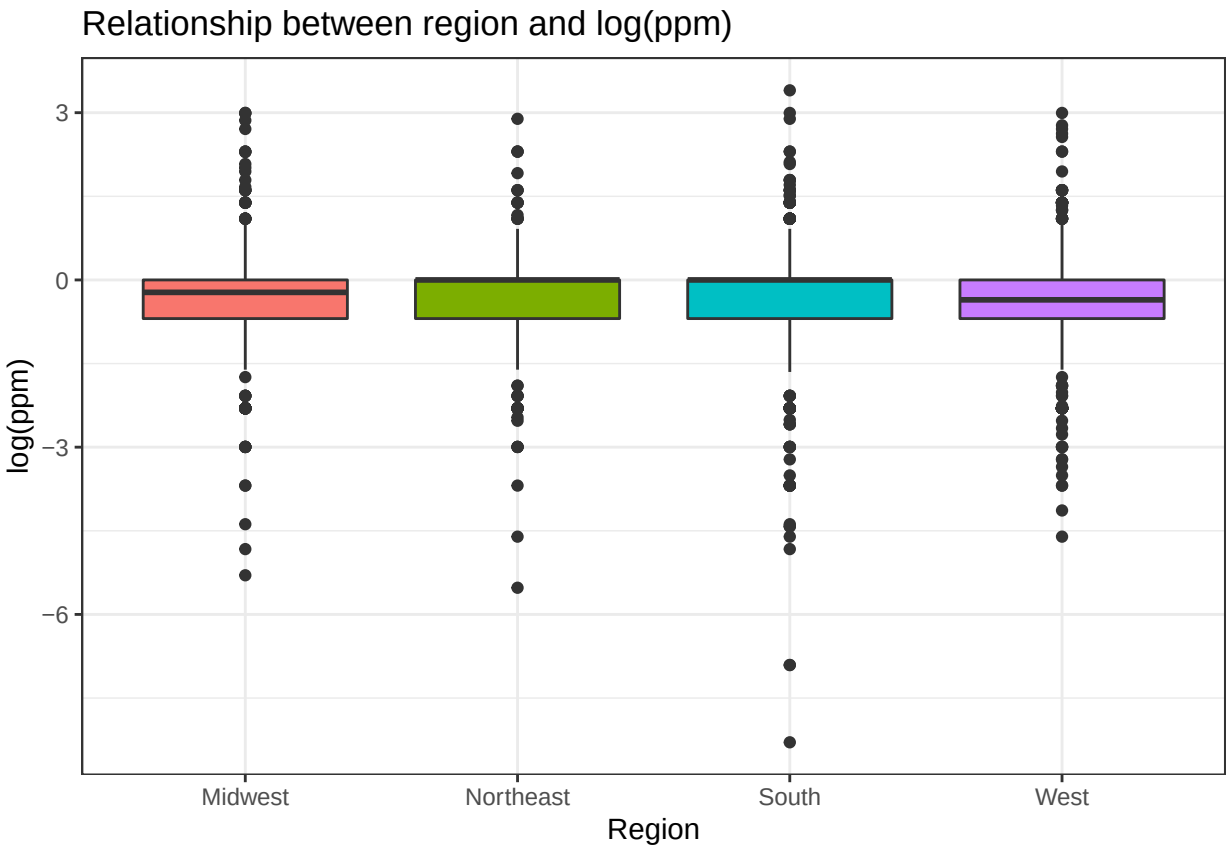
Dates


3
0
−3
−6
log(ppm)
2000
2005
2010
2015
2020
year


3
0
−3
−6
log(ppm)
0
2000
4000
6000
days



Geographical Information

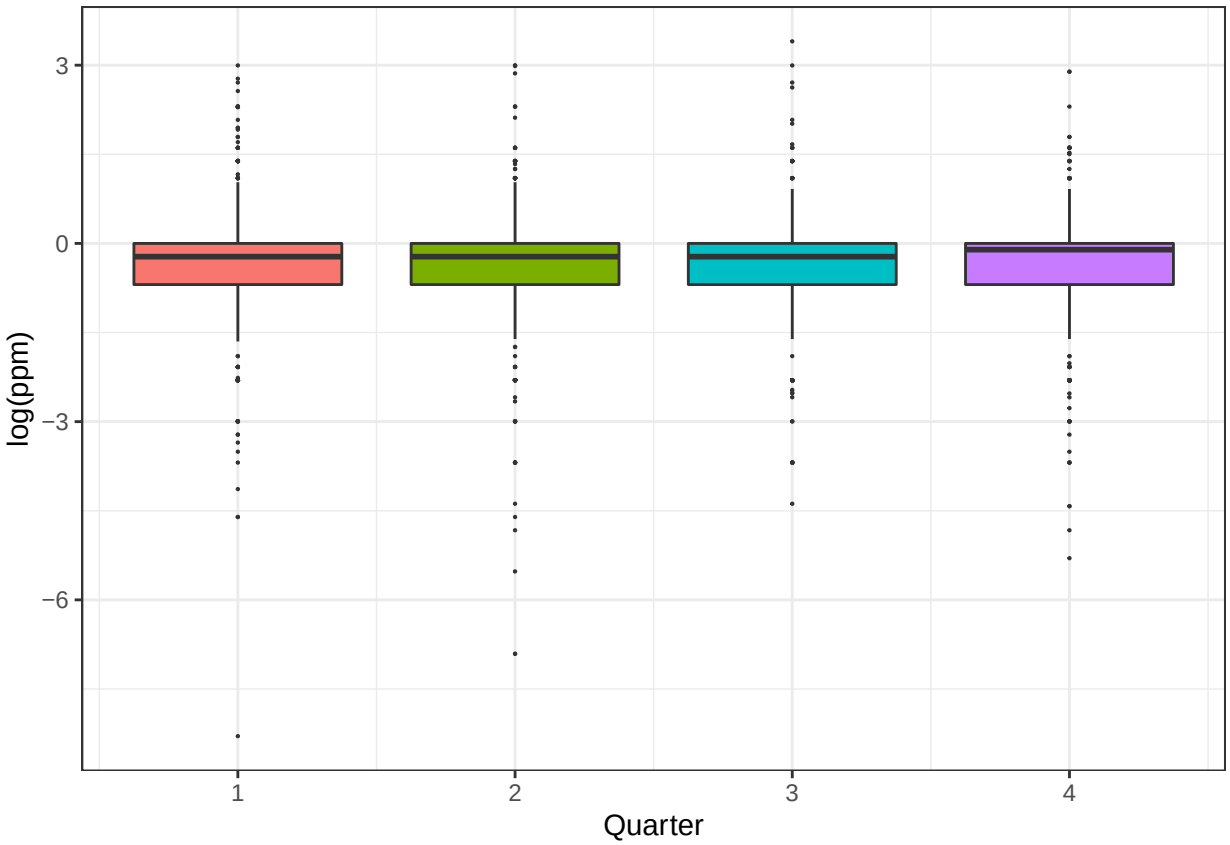


Form_temp

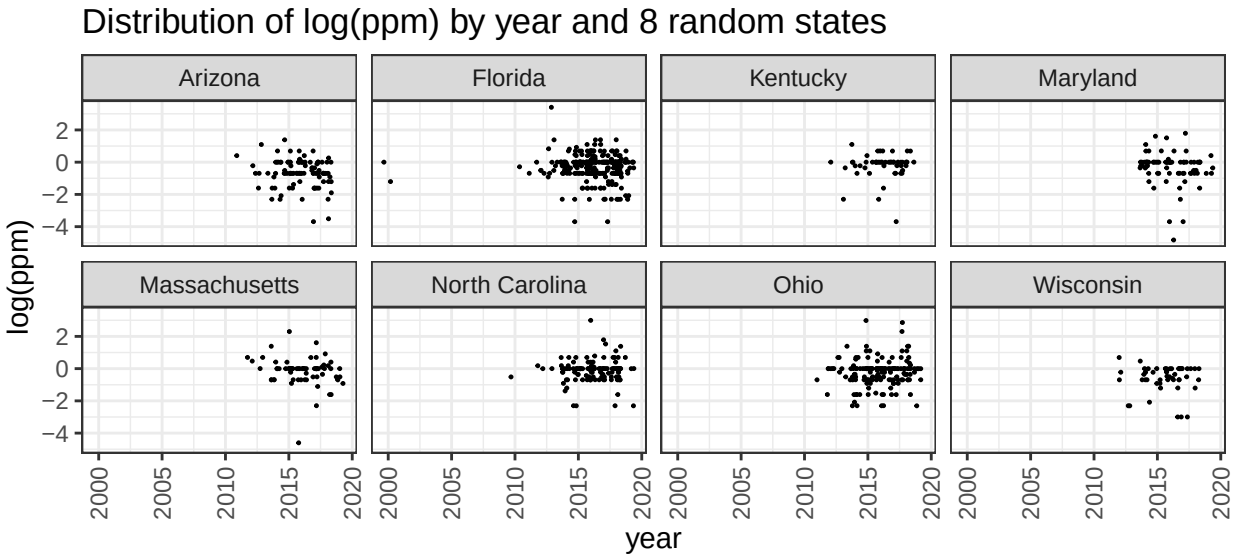
```
## # A tibble: 1 x 2
##   form_temp       n
##   <fct>       <int>
## 1 pill/tablet  4161
```

```
## # A tibble: 1 x 3
## # Groups:   form_temp [1]
##   form_temp   `is.na(mgstr)`     n
##   <fct>       <lgl>          <int>
## 1 pill/tablet FALSE           4161
```

Quarter

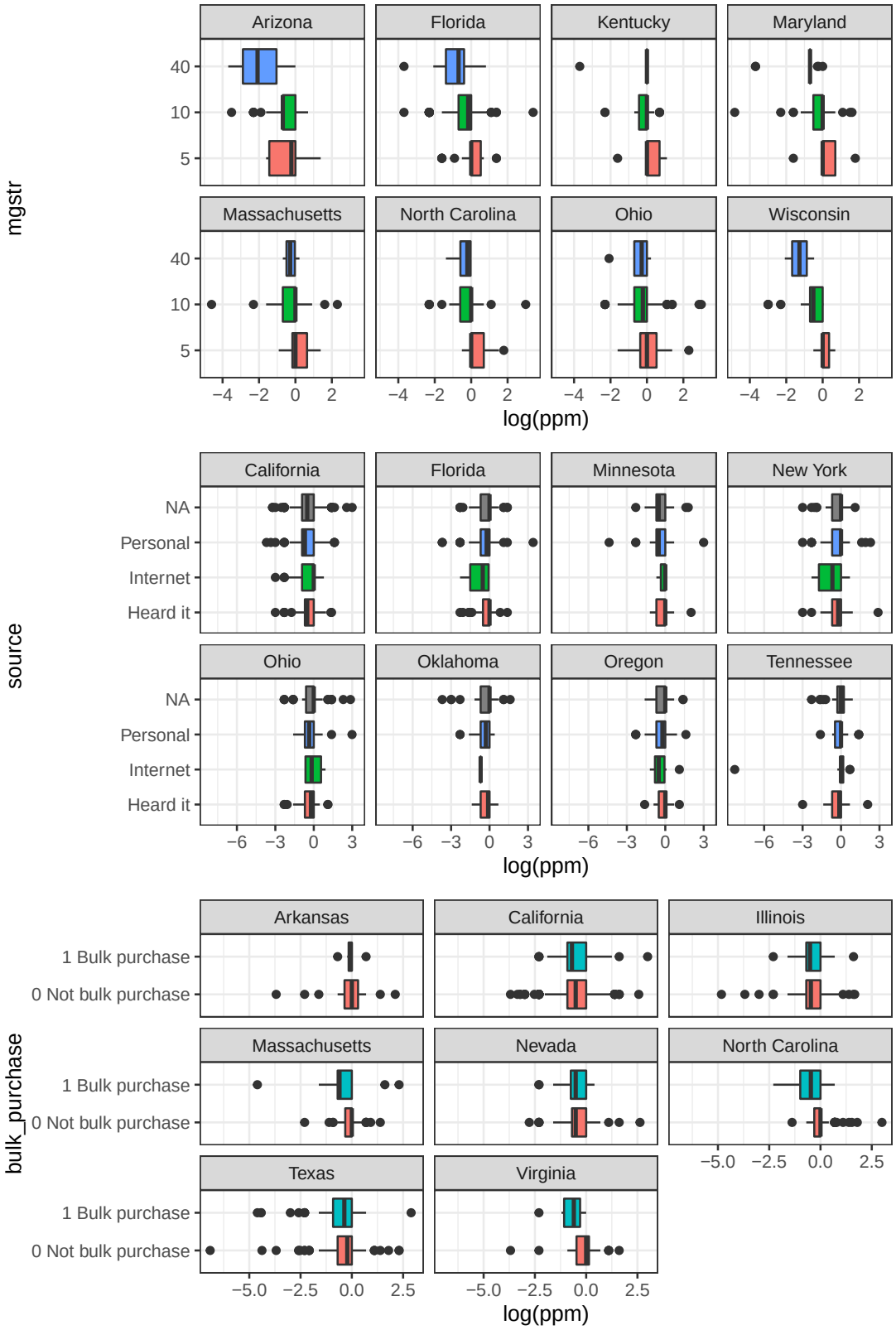


Assessment of Random Slopes



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

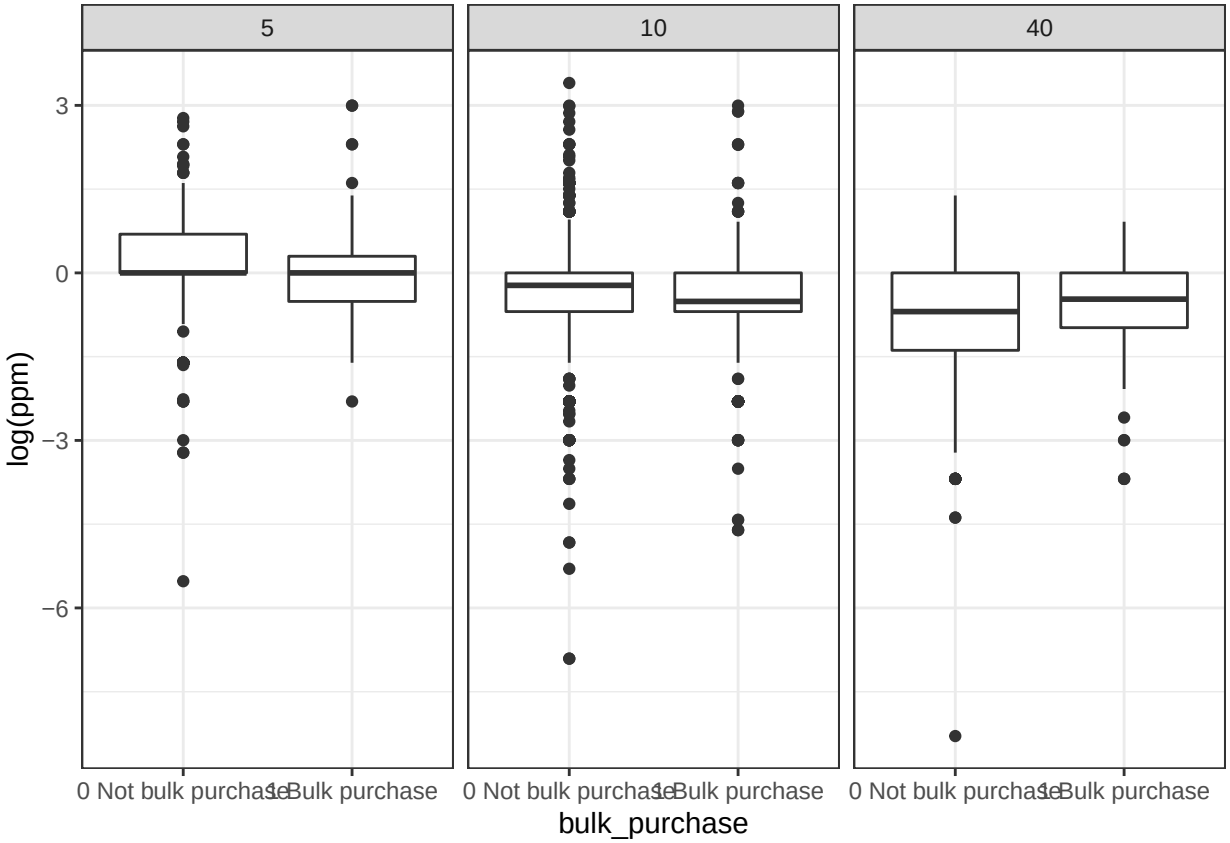
We do not observe difference in levels by state



Interaction plots

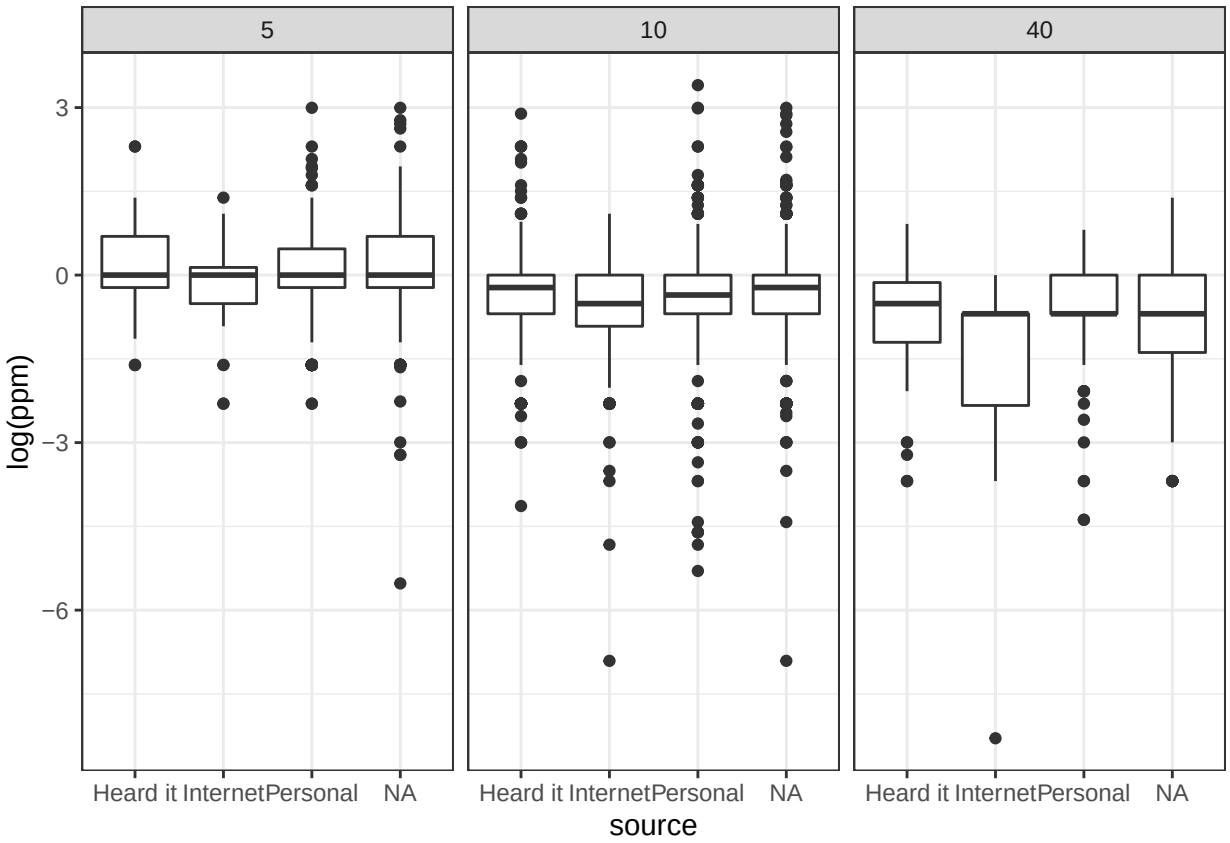
1.) `bulk_purchase` and `mgstr`

The following boxplot does reveal some slight variance in the effect of `bulk_purchace` on `log(ppm)` across values of `mgstr`



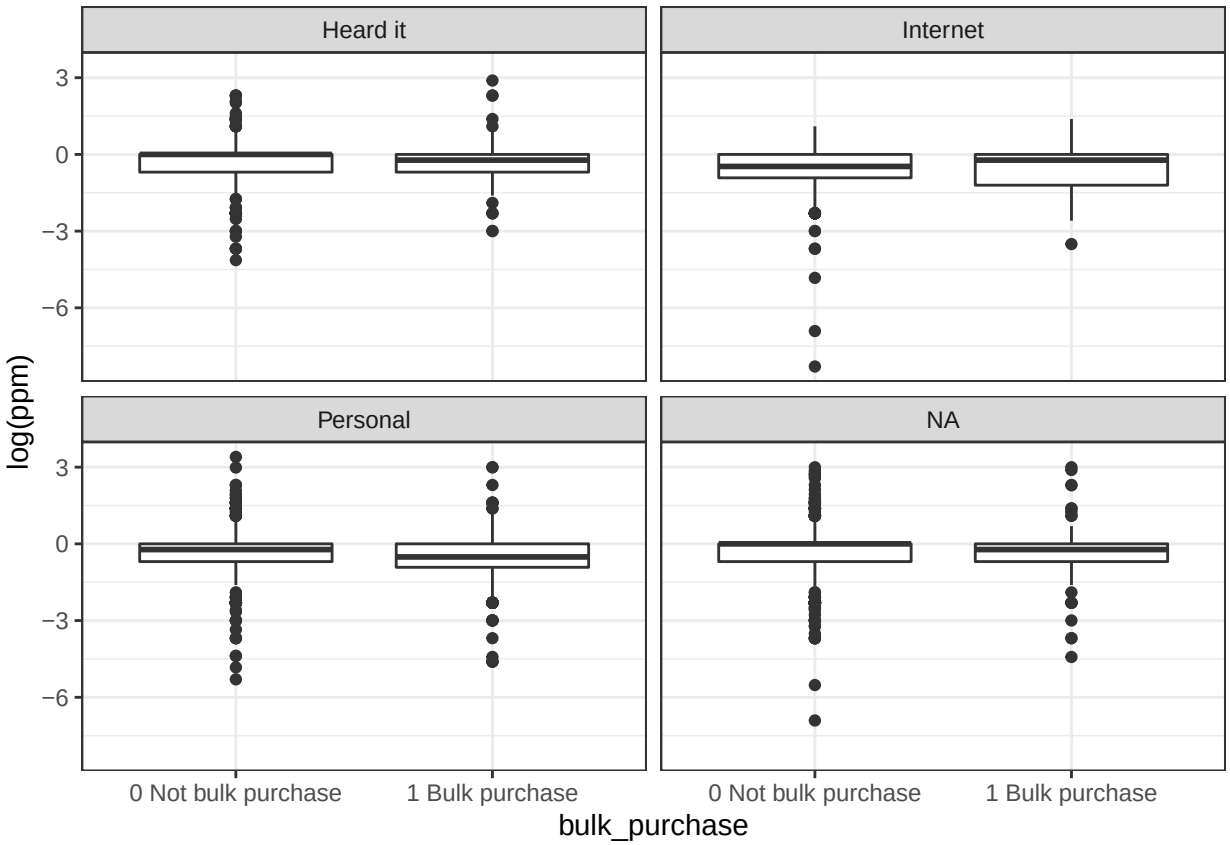
2.) `source` and `mgstr`

The following boxplot does reveal some slight variance in the effect of `source` on `log(ppm)` across different values of `mgstr`, though some of this effect could be distorted by lack of observations within certain categories.



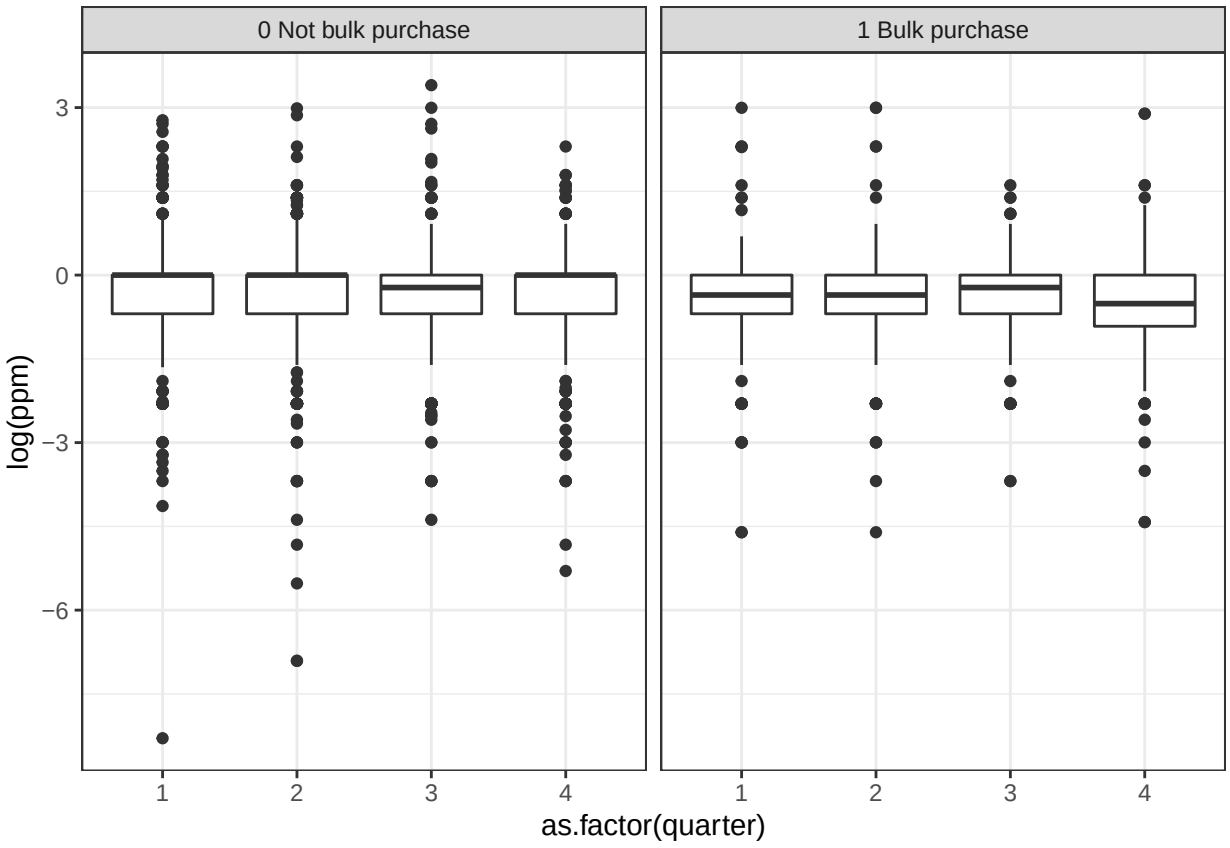
3.) `source` and `bulk_purchase`

The follow boxplot does not show substantial evidence that the relationship between log(ppm) and bulk_purchace varies across source.



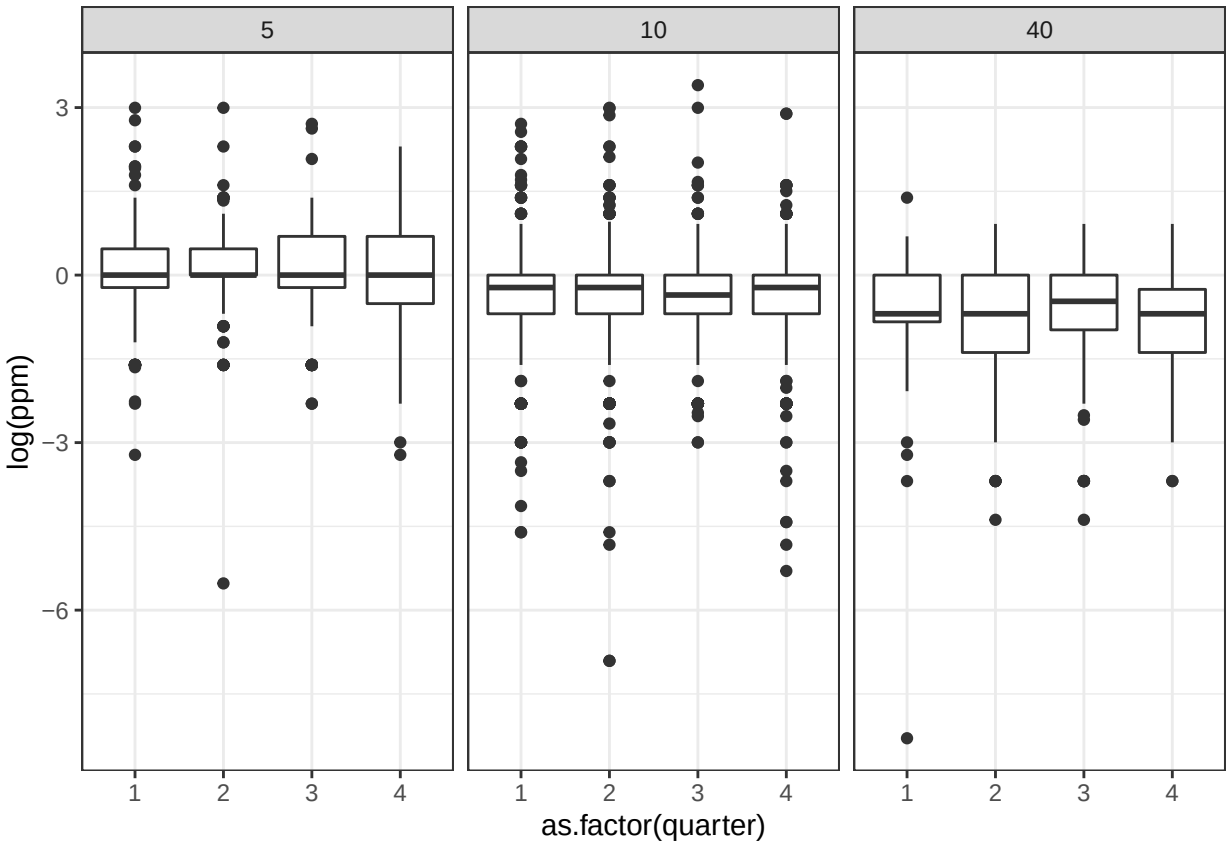
4.) `bulk_purchase` and `quarter`` ######

From the boxplot it appears that the relationship between quarter and log(ppm)changes slightly for different values of bulk_purchase.



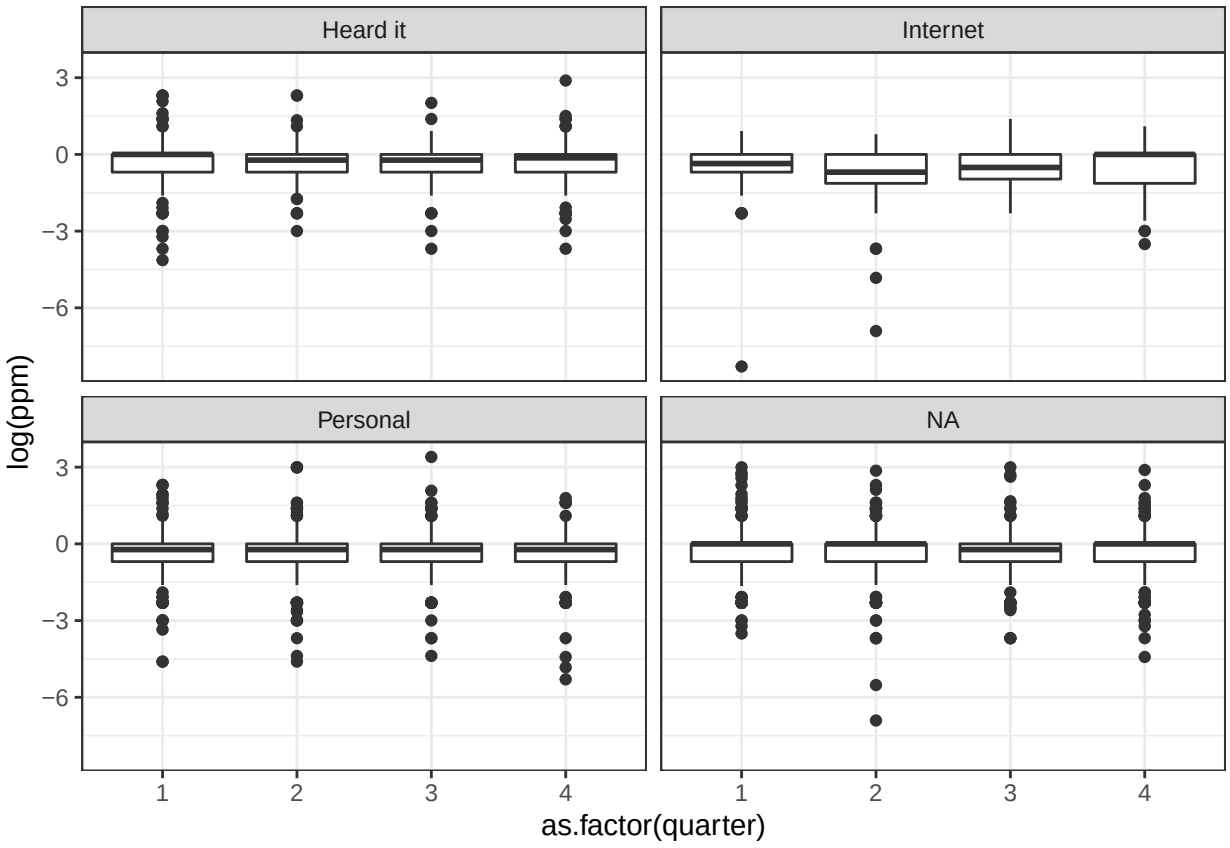
5.) `mgstr` and `quarter`` `#####`

There is some evidence that the effect of quarter on log(ppm) changes with thedosage unit.



6.) `source` and `quarter`

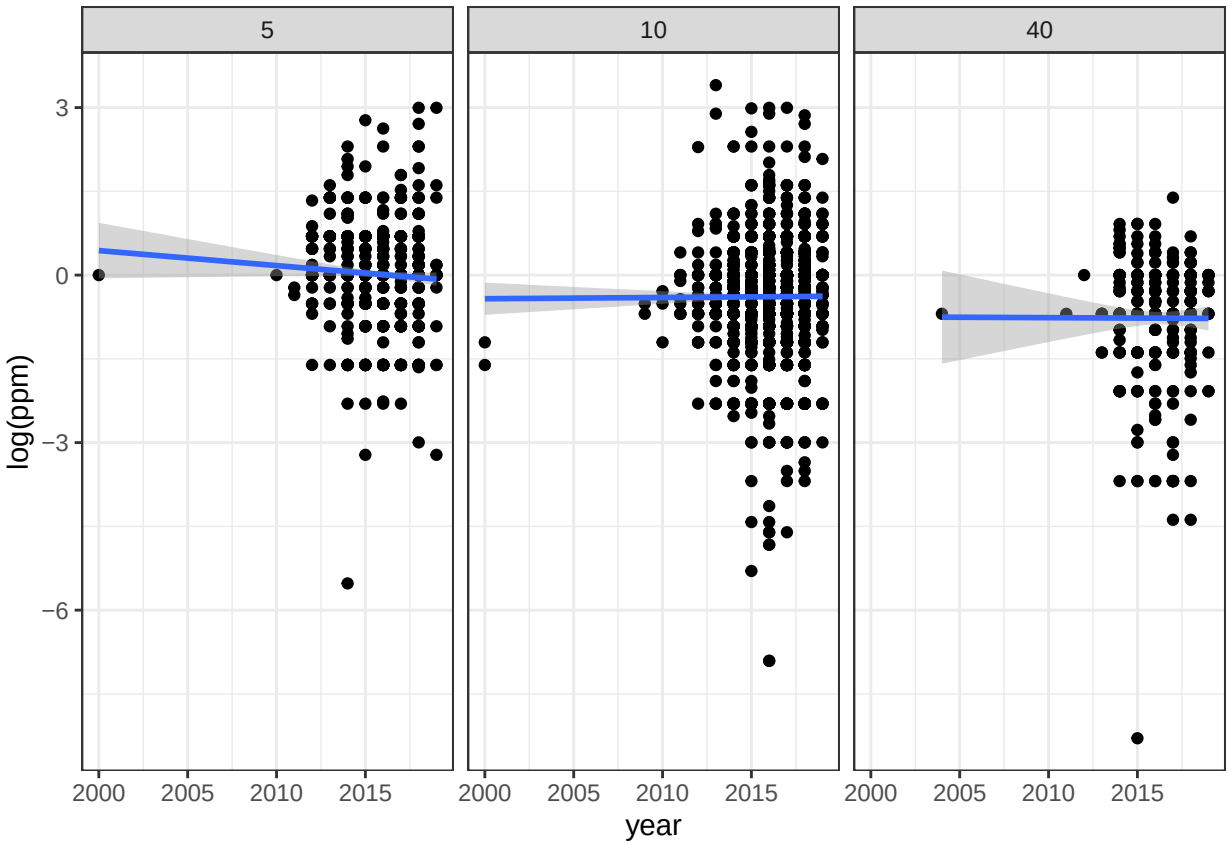
We do not see strong evidence that the relationship between quarter and log(ppm) changes across source.

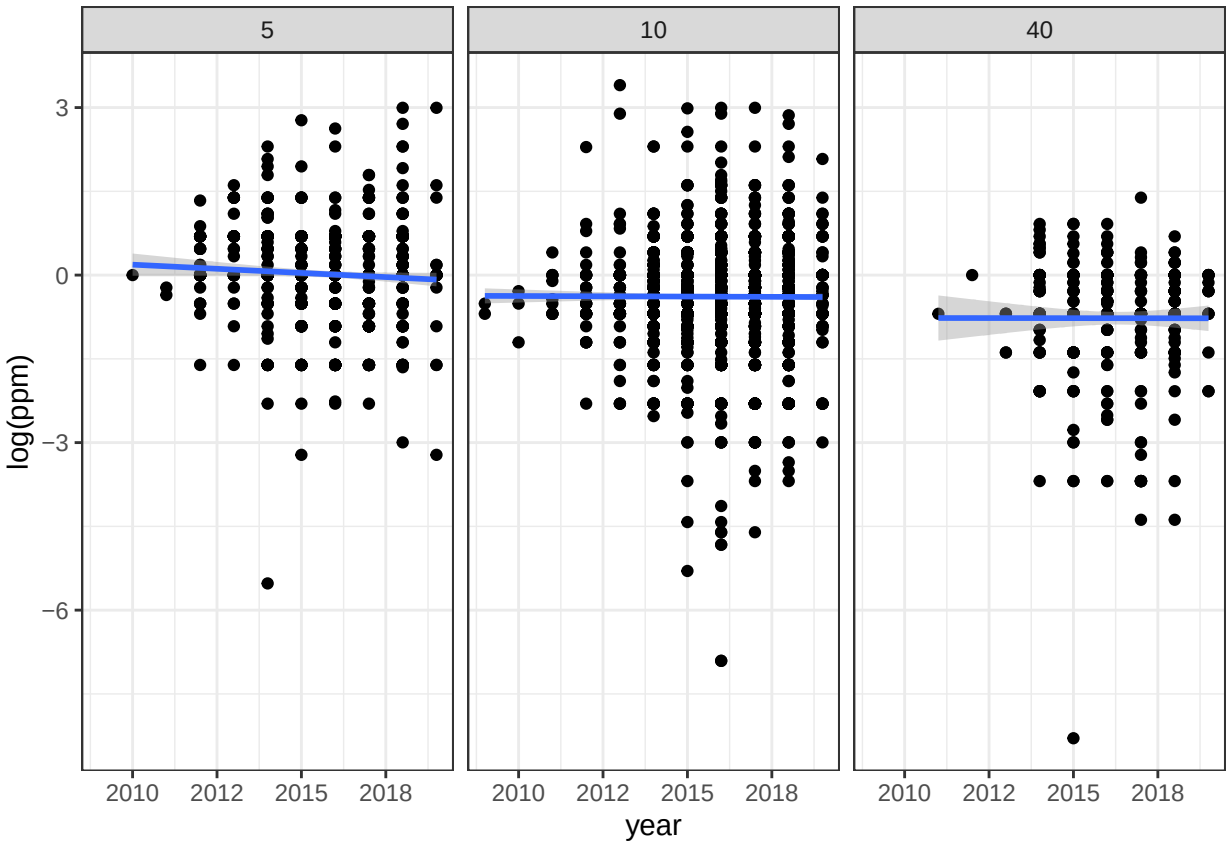


7.) Interactions between factor variables and year

7.1) `mgstr`

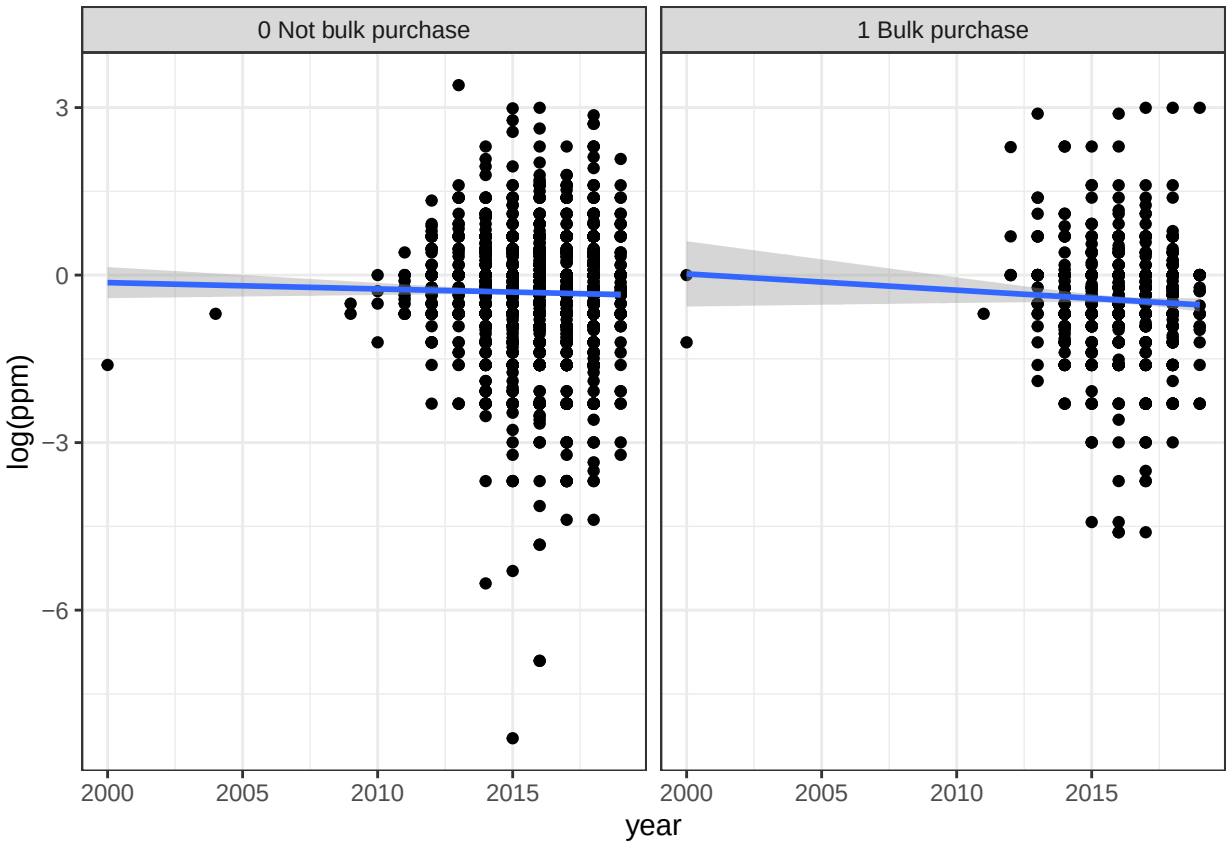
There is not strong evidence that `mgstr` effects the relationship between `year` and `log(ppm)`.

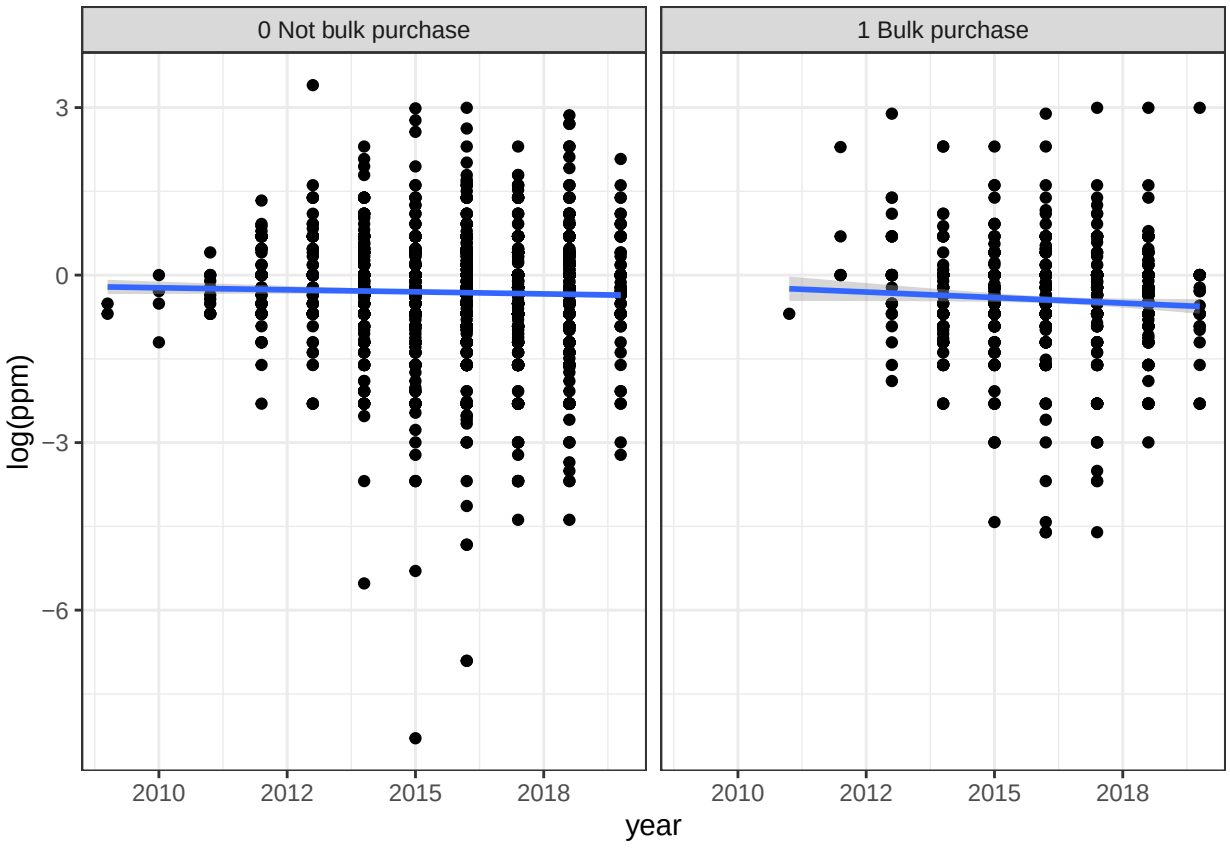

5
10
40
3
0
−3
−6
log(ppm)
2000
2005
2010
2015
year



7.2) `bulk_purchase`

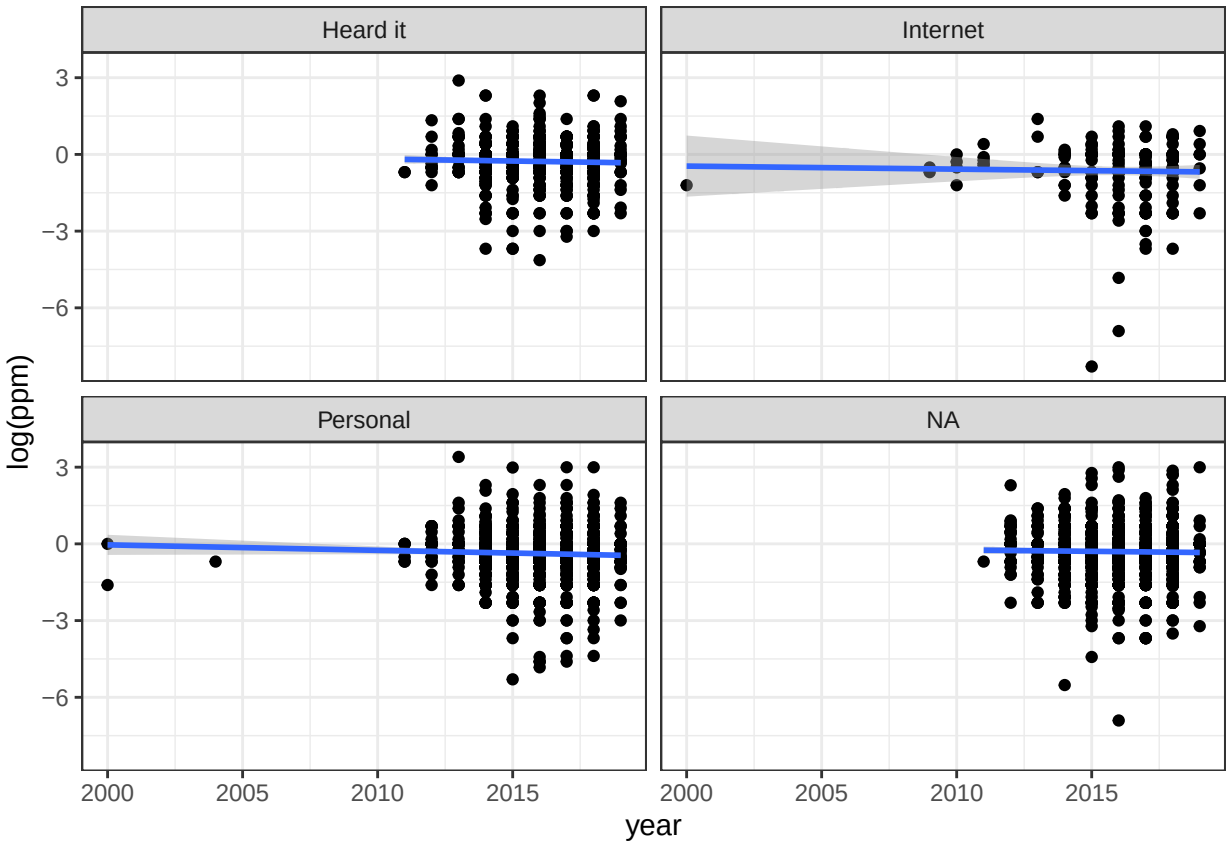
There is not strong evidence that `bulk_purchase` effects the relationship between `year` and `log(ppm)`.

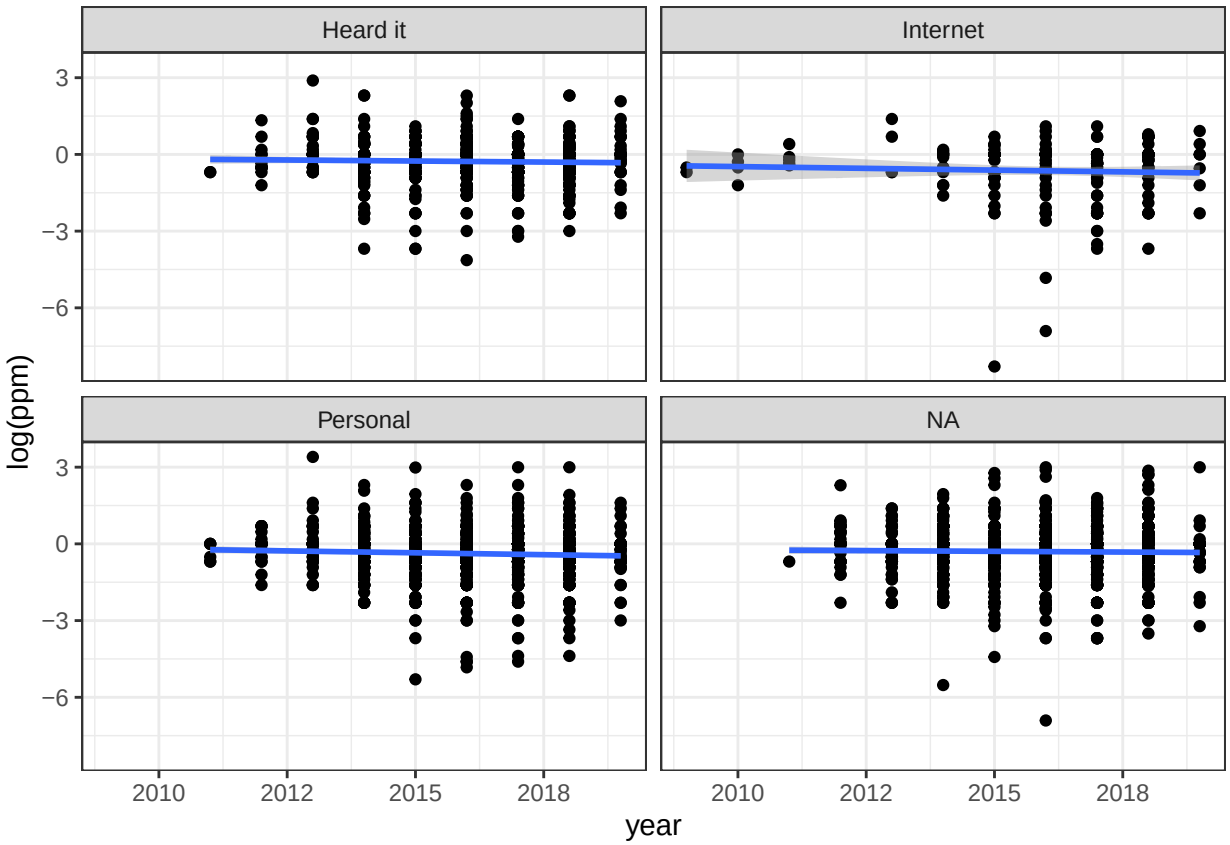

0 Not bulk purchase
1 Bulk purchase
log(ppm)
3
0
−3
−6
2000
2005
2010
2015
year



7.3) `source`

There is not strong evidence that `source` effects the relationship between `year` and `log(ppm)`.


Heard it
Internet
Personal
NA
log(ppm)
3
0
−3
−6
2000
2005
2010
2015
year



Outside of `mgstr` and `quarter`, `bulk_purchase` and `quarter`, `source` and, and `mgstr` `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.

Frequentist Model Output

Fitting the model with frequentist MLE, we have the following results.

Bayesian Model Comparison

Model Fitting

Here we imposed a prior on the parameters and fit the model using a Bayesian approach to see if we had differing estimates using this approach. Since we did not have much information about the model, we used non-informative priors:

$$\beta_j \overset{iid}{\sim} \text{Normal}(0, 1)$$
$$\tau^2 \sim \text{InvGamma}(0.1, 0.1)$$
$$\sigma^2 \sim \text{InvGamma}(0.1, 0.1)$$

Estimates Comparison

Table 7: Fixed effect estimates comparison

	Frequentist Model Estimate	Bayesian Model Esimate
(Intercept)	0.1891	0.1883
mgstr10	-0.4345	-0.4329
mgstr40	-0.8240	-0.8213
bulk_purchase1 Bulk purchase	-0.1325	-0.1327
sourceInternet	-0.4013	-0.4017
sourcePersonal	-0.1051	-0.1049

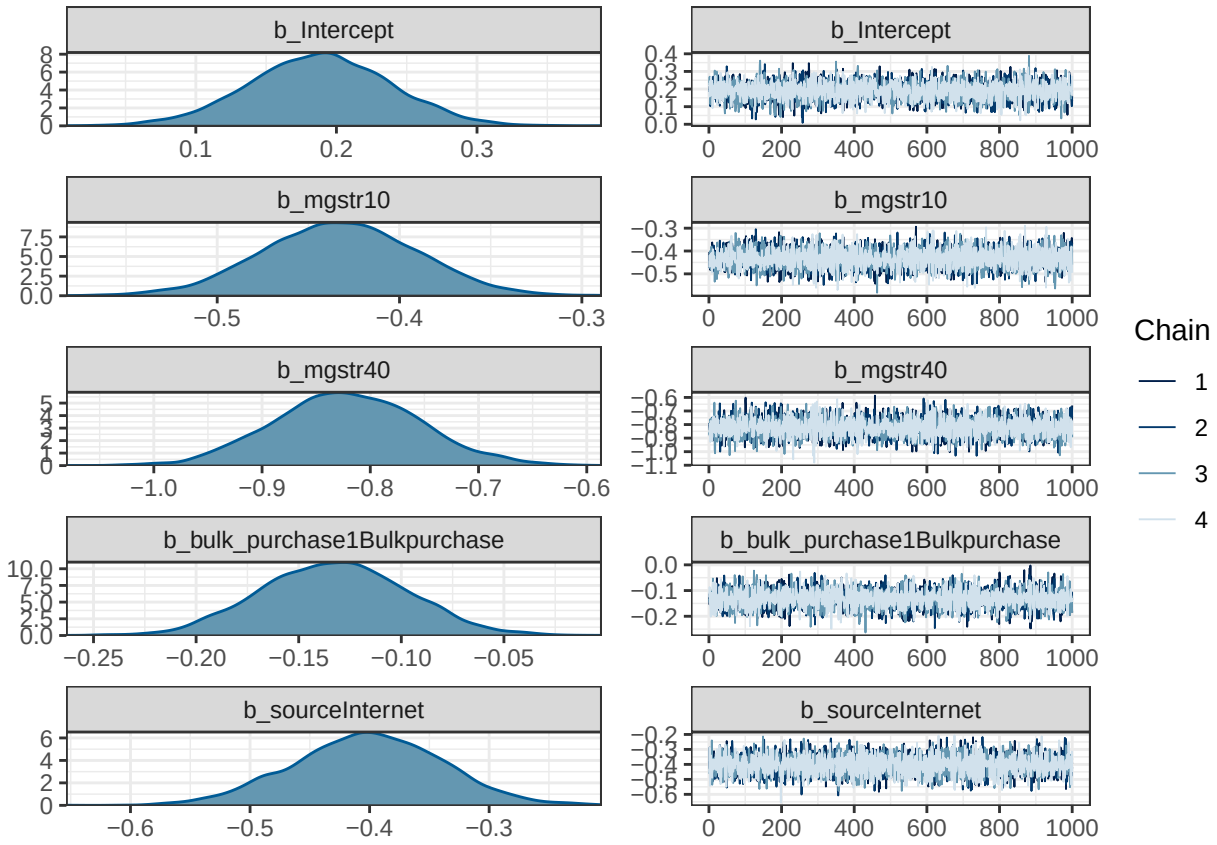
Table 8: Variance estimates comparison

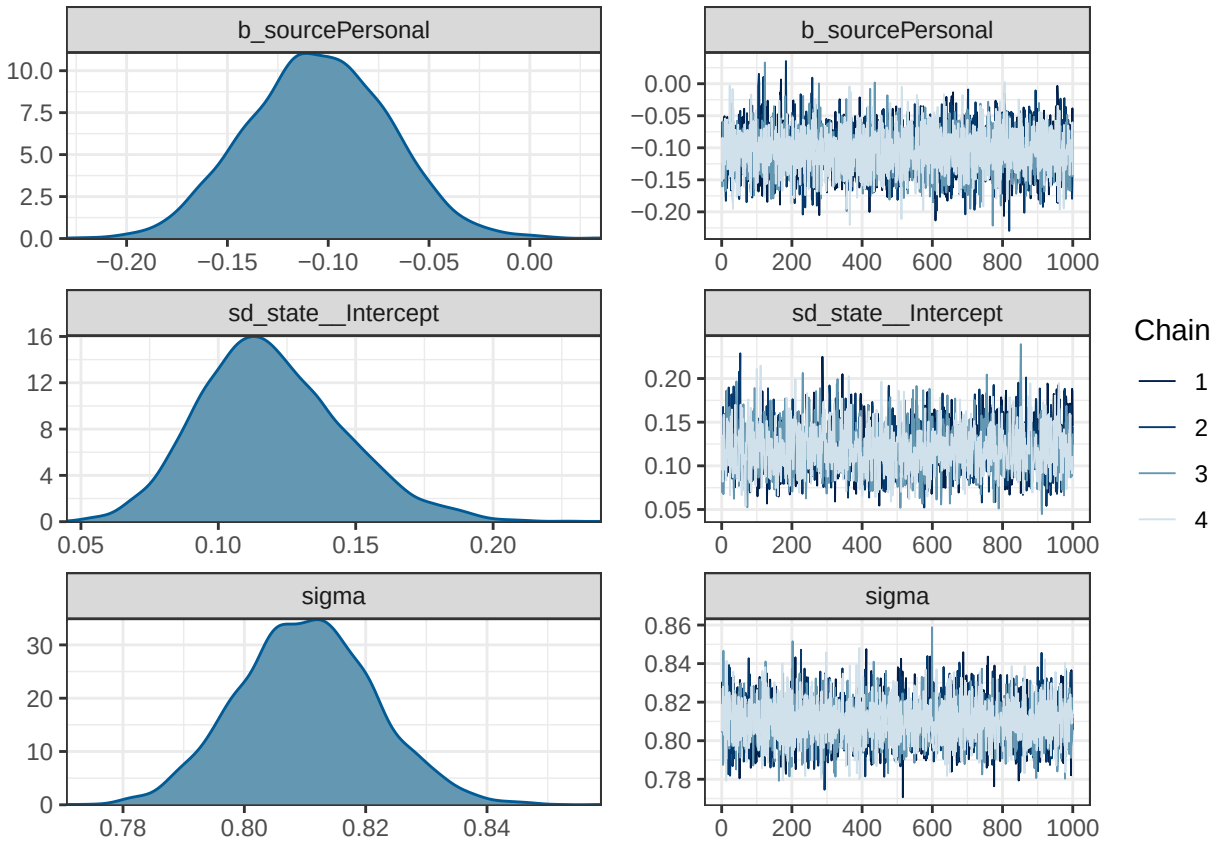
Source of Variation	Bayesian Model Estimate	Frequentist Model Estimate
Residual	0.6566	0.6548
State	0.0143	0.0124

From the above tables, we can see that the results given by Bayesian setting is almost the same as that from the frequentist setting.

Posterior Checks

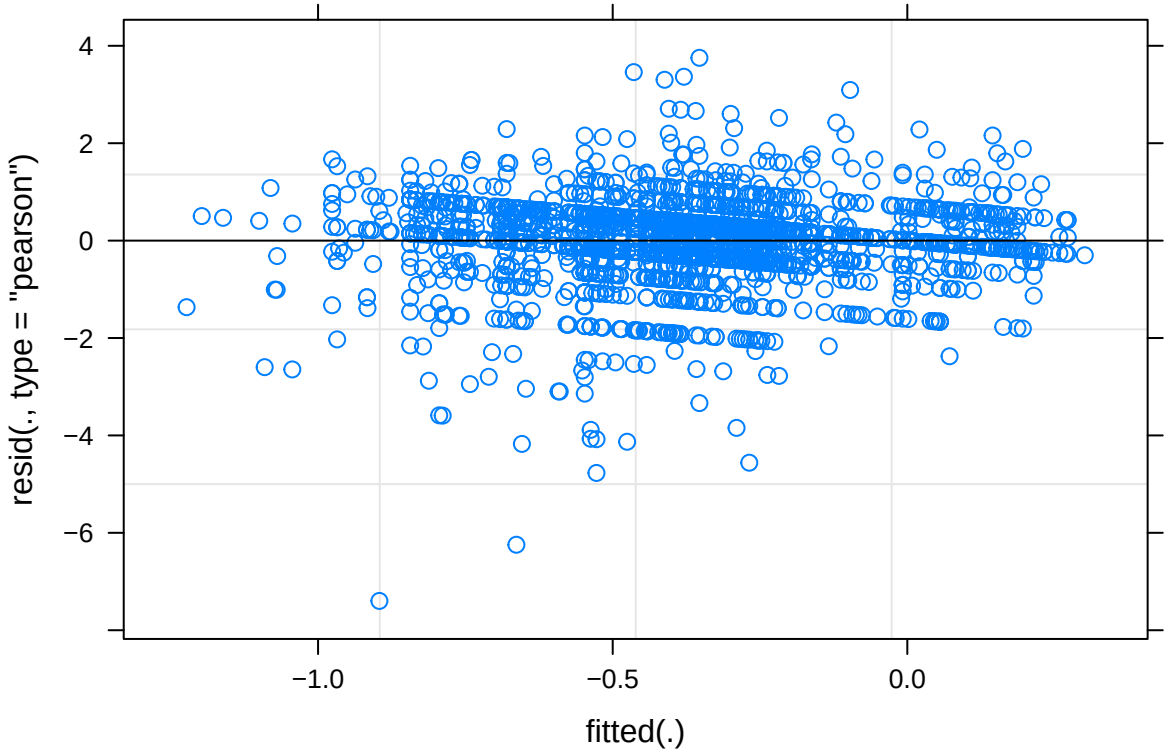
The below plots show our posterior distributions for all parameters of interest and traceplots to check that the sampling chains converged, which they did as shown below.

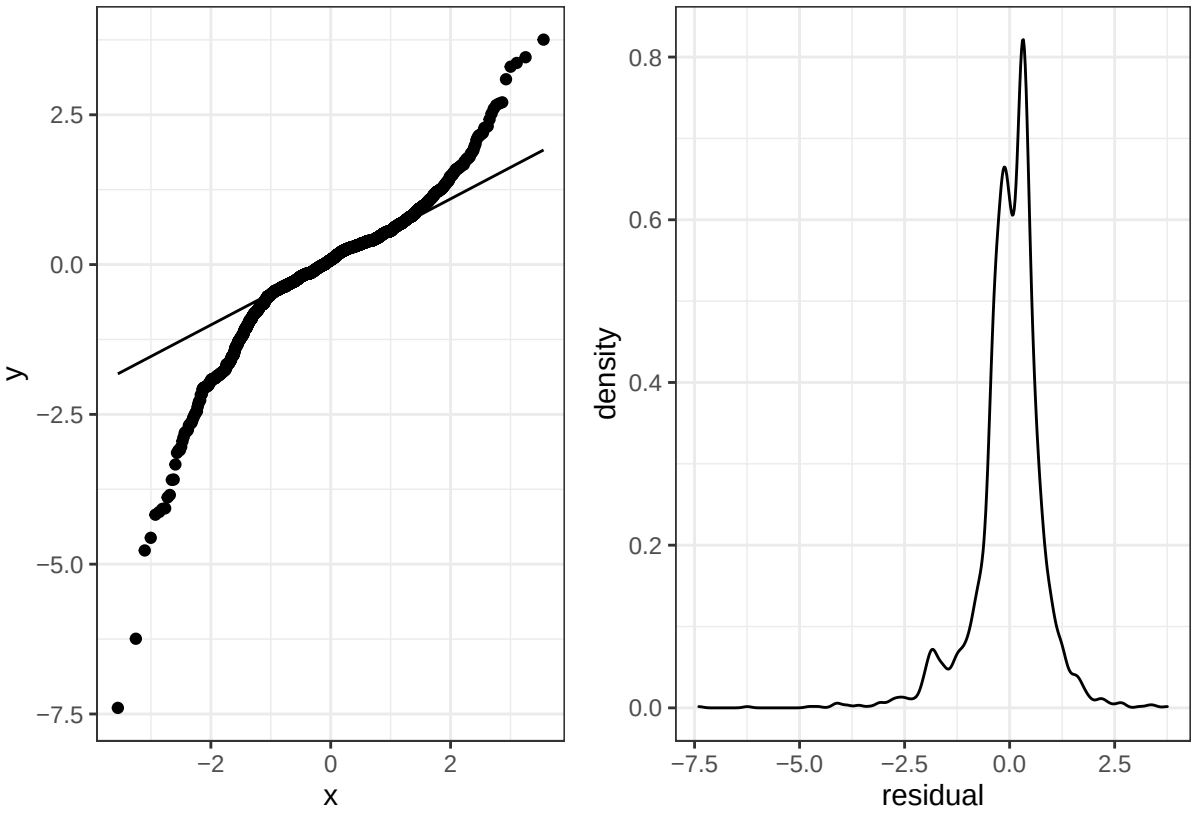

b_Intercept
b_mgstr10
b_mgstr40
b_bulk_purchase1Bulkpurchase
b_sourceInternet
Chain
1
2
3
4


b_sourcePersonal
10.0
7.5
5.0
2.5
0.0
−0.20
−0.15
−0.10
−0.05
0.00
b_sourcePersonal
0.00
−0.05
−0.10
−0.15
−0.20
0
200
400
600
800
1000
sd_state__Intercept
16
12
8
4
0
0.05
0.10
0.15
0.20
sd_state__Intercept
0.20
0.15
0.10
0.05
0
200
400
600
800
1000
Chain
1
2
3
4
sigma
30
20
10
0
0.78
0.80
0.82
0.84
sigma
0.86
0.84
0.82
0.80
0.78
0
200
400
600
800
1000

Model Diagnostics

Residual Plots



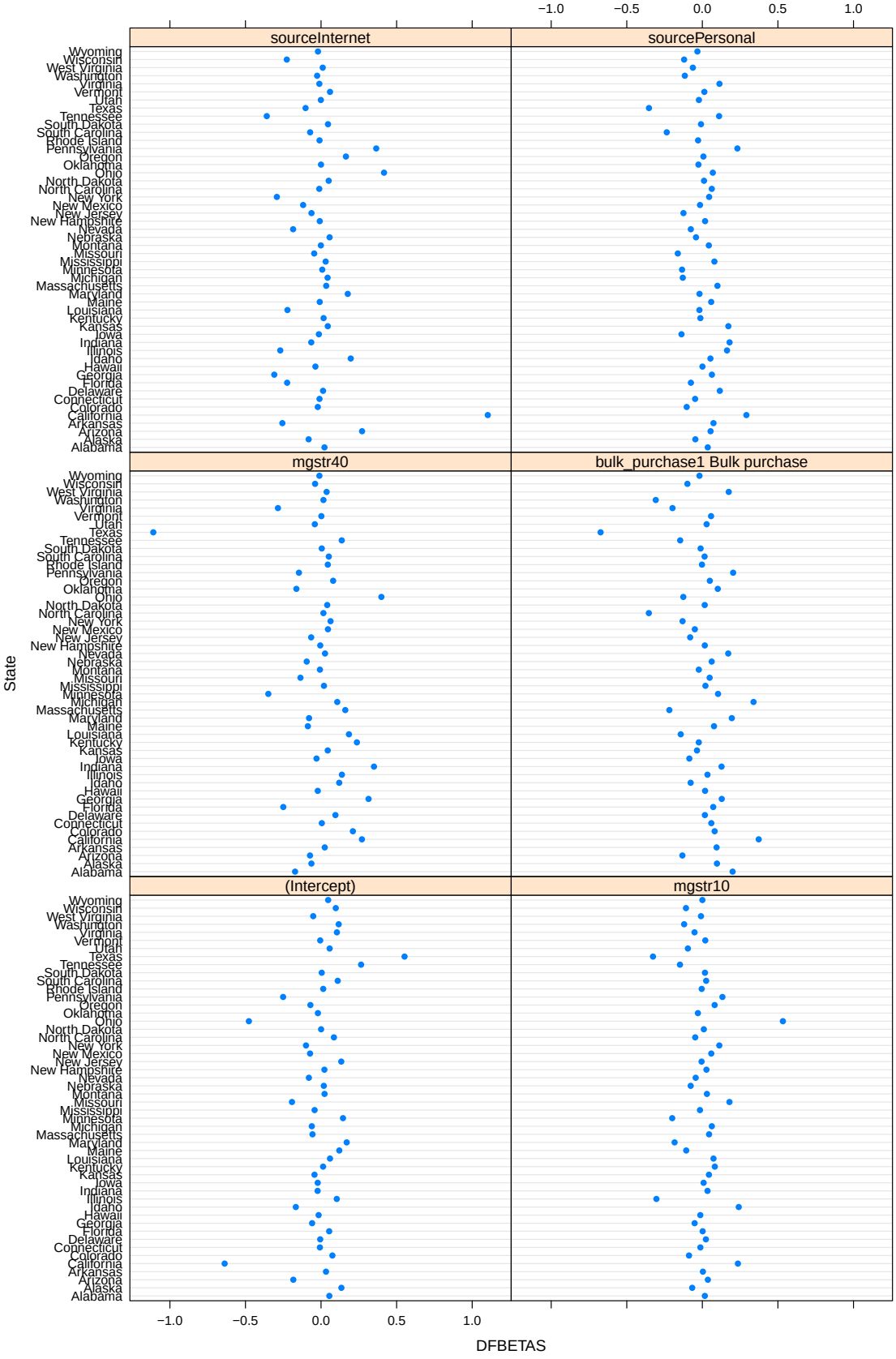


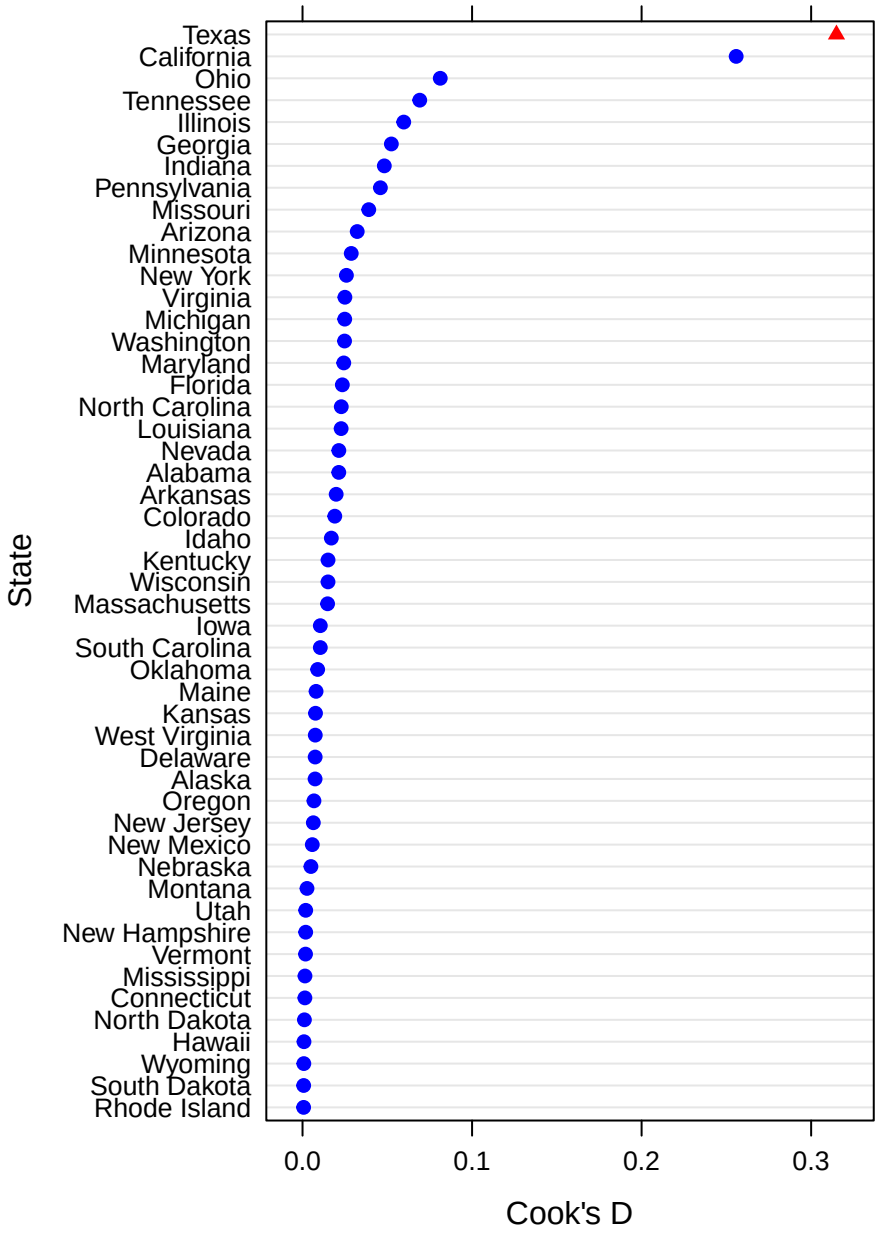
Influential Groups

Table of DFBETAS

Table 9: Level of influence 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


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


Texas
California
Ohio
Tennessee
Illinois
Georgia
Indiana
Pennsylvania
Missouri
Arizona
Minnesota
New York
Virginia
Michigan
Washington
Maryland
Florida
North Carolina
Louisiana
Nevada
Alabama
Arkansas
Colorado
Idaho
Kentucky
Wisconsin
Massachusetts
Iowa
South Carolina
Oklahoma
Maine
Kansas
West Virginia
Delaware
Alaska
Oregon
New Jersey
New Mexico
Nebraska
Montana
Utah
New Hampshire
Vermont
Mississippi
Connecticut
North Dakota
Hawaii
Wyoming
South Dakota
Rhode Island
State
0.0
0.1
0.2
0.3
Cook's D