STA610 Case Study 1

Emily Gentles, Weiyi Liu, Jack McCarthy, Qinzhe Wang

11 October, 2021

Qinzhe - Coordinator & Checker: Double-checks the work for reproducibility and errors. Also responsible for submitting the report and presentation files. Coordinator: Keeps everyone on task and makes sure everyone is involved. Also responsible for coordinating team meetings and defining the objectives for each meeting.

Emily - Presenter: Primarily responsible for organizing and putting the team presentations together.

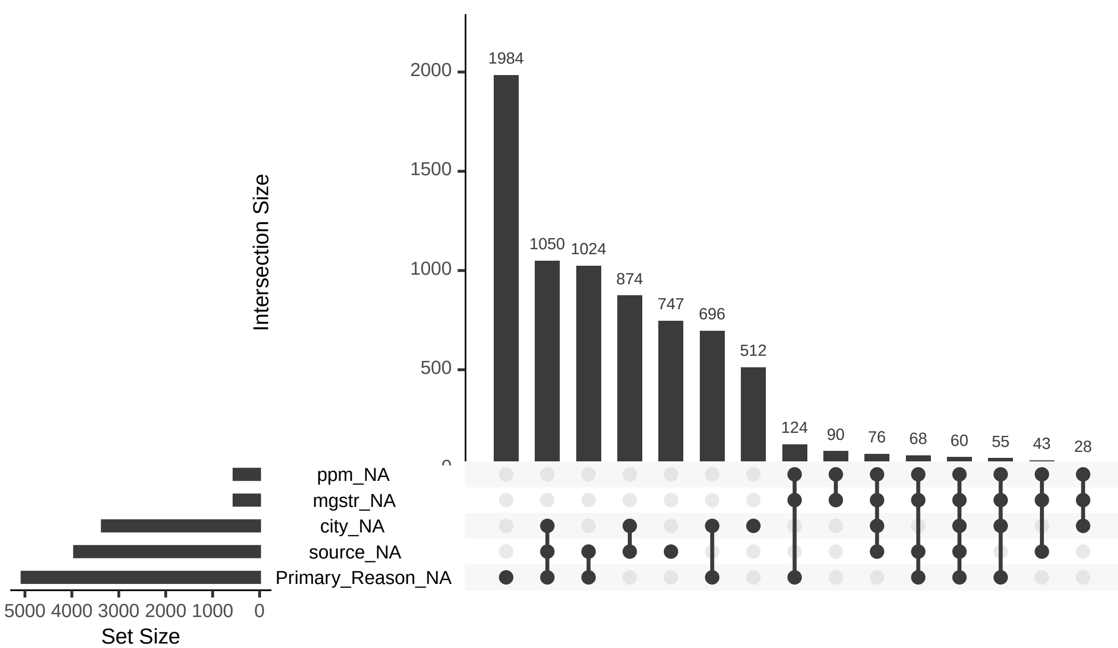
Jack - Programmer: Primarily responsible for all things coding. The programmer is responsible for putting everyone's code together and making sure the final product is "readable".

Weiyi - Writer: Primarily responsible for putting together the final report.

Introduction

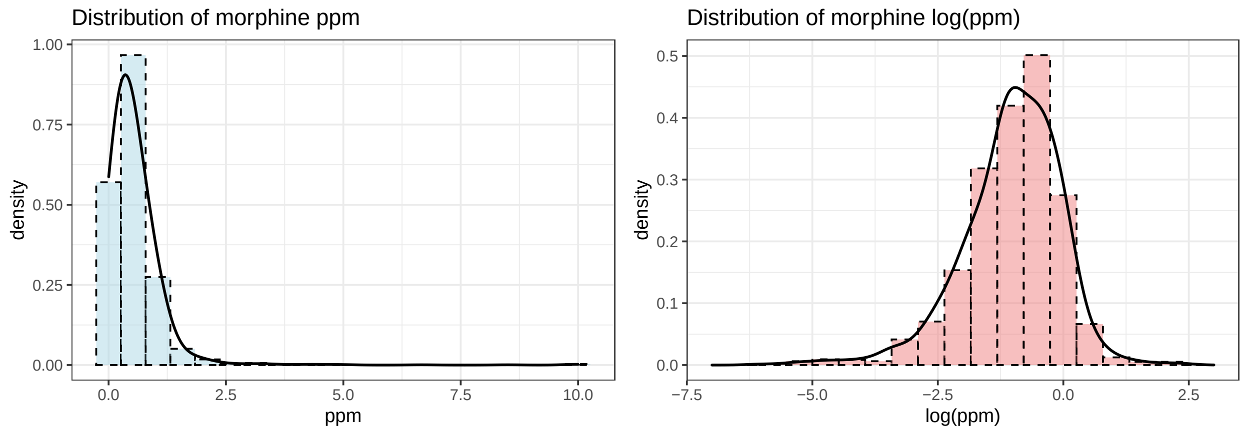
EDA

Missing Values



Response Distribution

First, a look at the distributions of the response variable "ppm". Observations with ppm between the 0.1 and 99.9 percentiles were considered so as to avoid the influence of extreme outliers on the analysis of the ppm distribution.



The distribution of ppm is clearly right-skewed, and it is strictly nonnegative in value, so a log transformation may be appropriate. The distribution of log(ppm) is given above, and appears closer to the desired normal.

state vs. log(ppm)

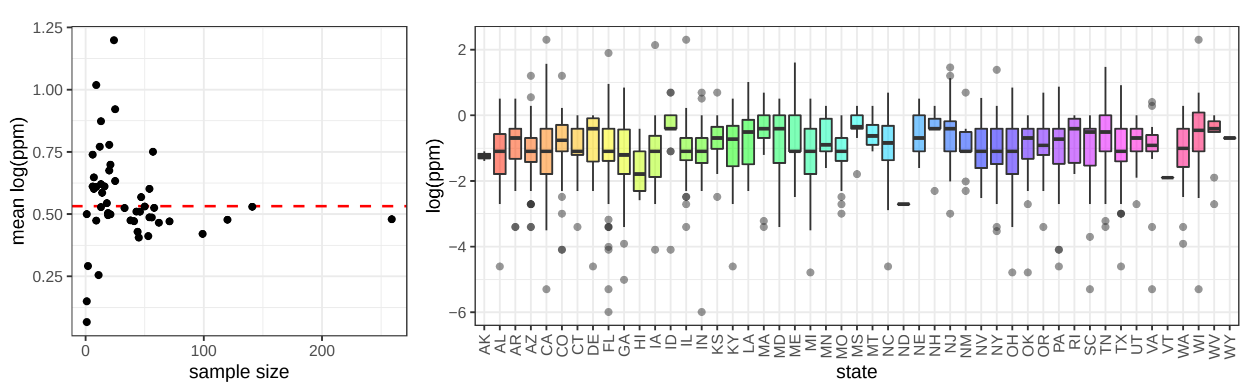
We see that there are 4 states that have a sample size of 1, North Dakota, Vermont, Washington DC, and Wyoming, as well as 1 state that has a sample size of 2, Alaska. Due to the extremely small sample sizes we decided to remove these states form our dataset to avoid computational instability.

Table 1: 7 States with Smallest Sample Size

North Dakota	Vermont	Washington, DC	Wyoming	Alaska	New Hampshire	Rhode Island
1	1	1	1	2	6	6

Table 2: 7 States with Largest Sample Size

Pennsylvania	Ohio	Arizona	Michigan	Texas	Florida	California
58	62	71	99	120	141	259

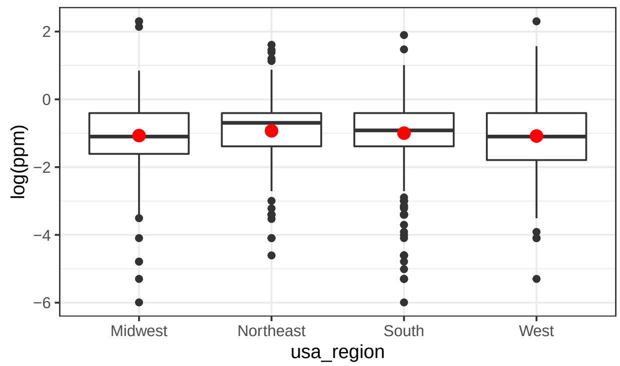


We observe that the within-state means for states with higher sample sizes in general adhere more closely to the grand mean. It is also evident that the log(ppm) distributions differ little as compared to the within-state variance. This is conducive to the borrowing of information between states.

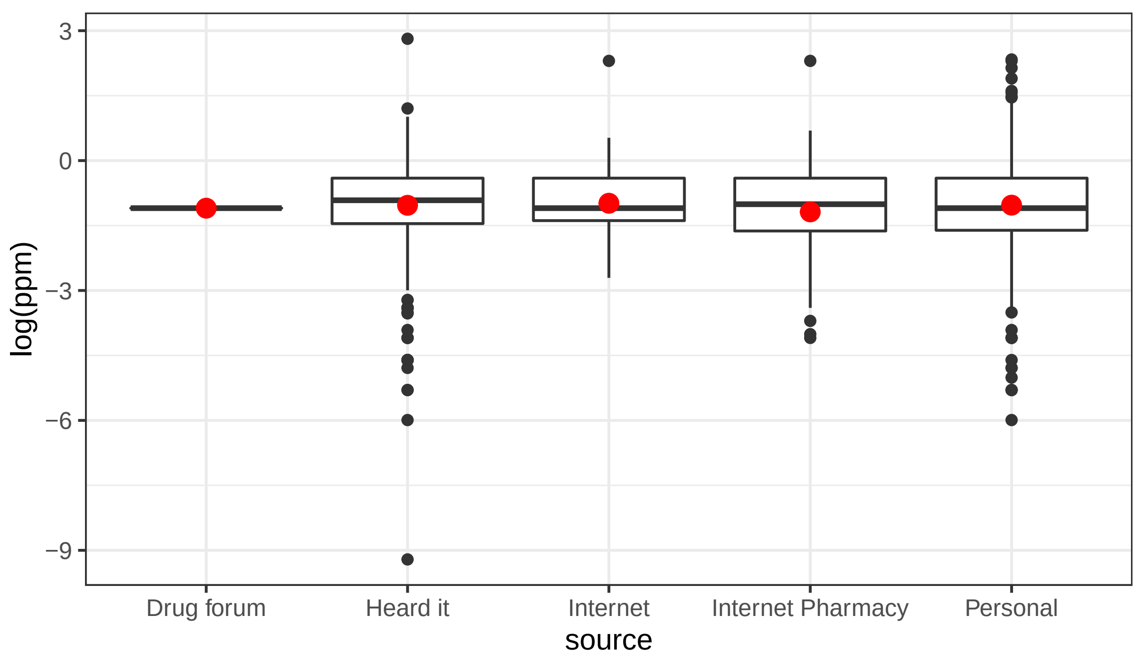
region vs. log(ppm)

We also have access to the broader region in which a purchase is made. This could be useful if we wanted to develop a simpler model that still captured variation by purchase location.

	usa_region	n	mean
1	Midwest	386	−1.069
2	Northeast	191	−0.930
3	South	673	−0.998
4	West	583	−1.083

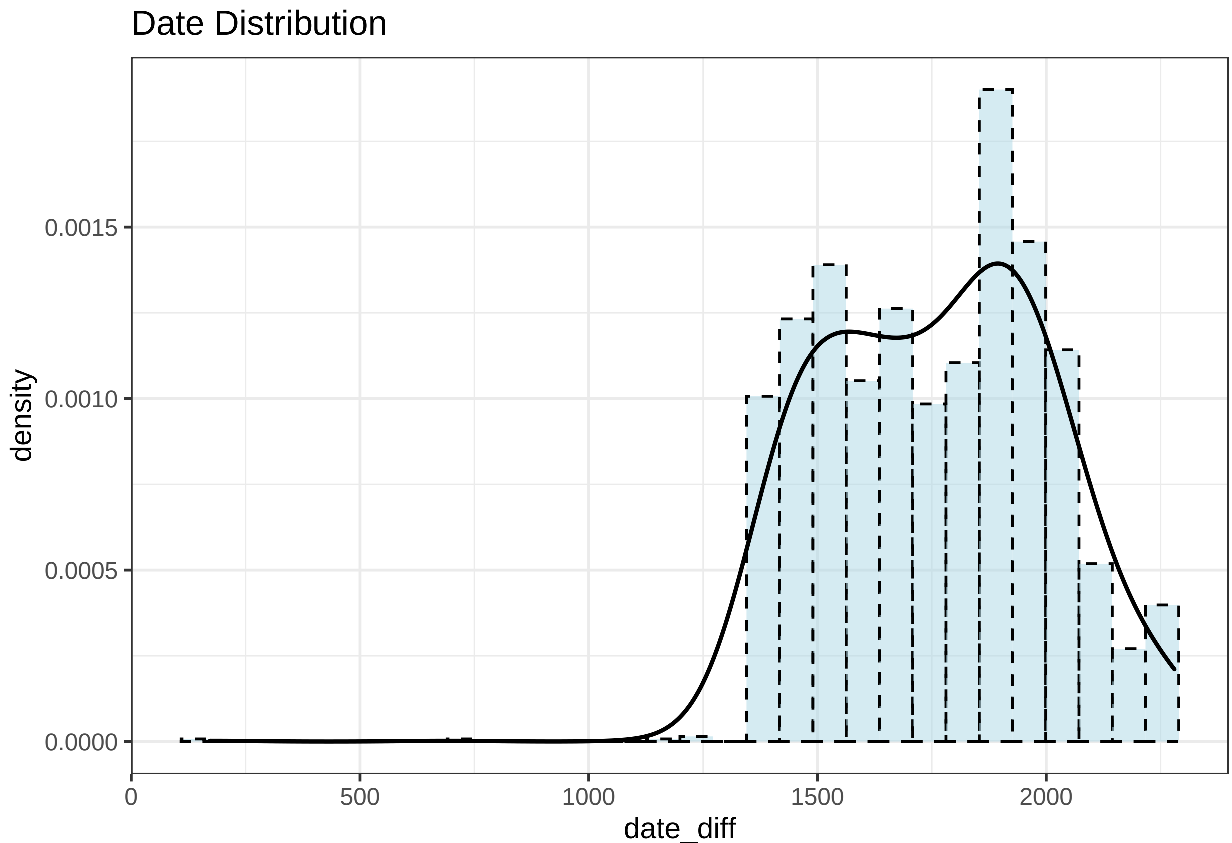


source vs. log(ppm)

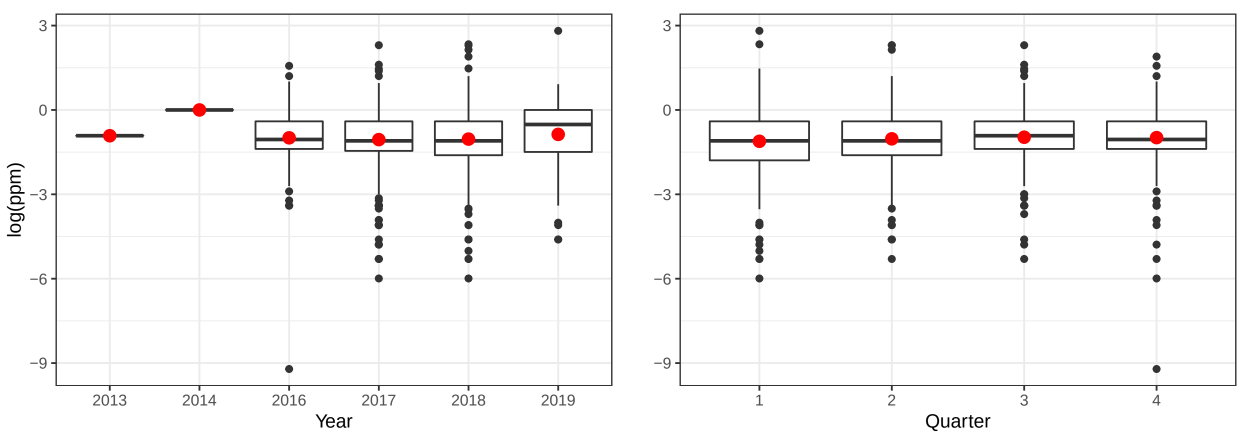


date

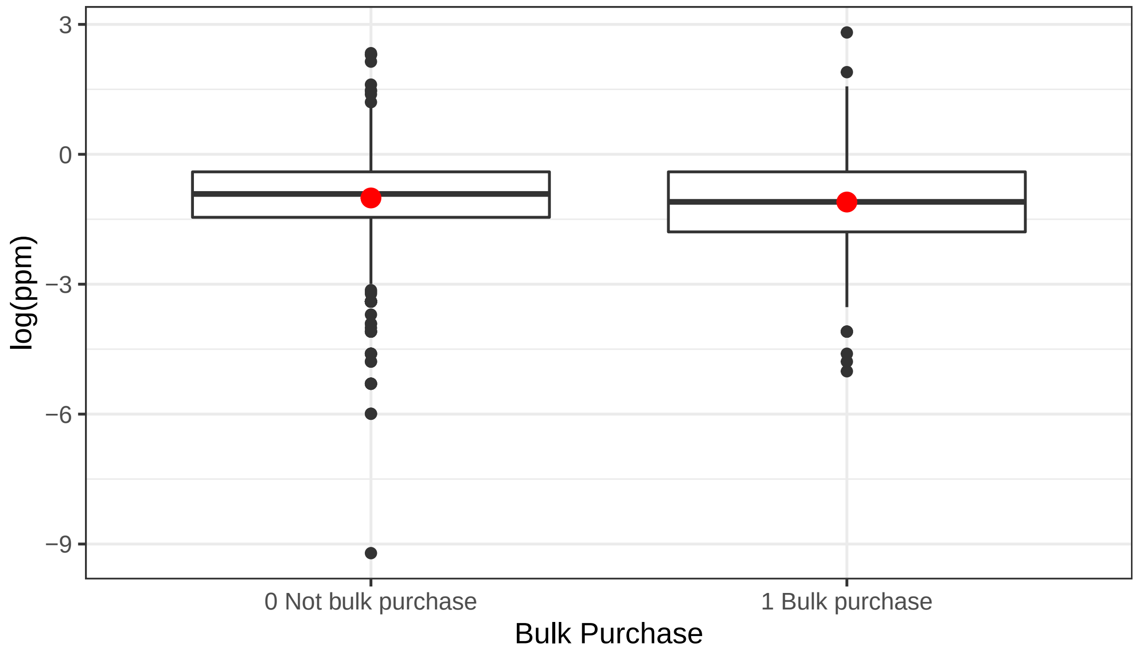
record `price_date` as a continuous variable counting days from some start date.



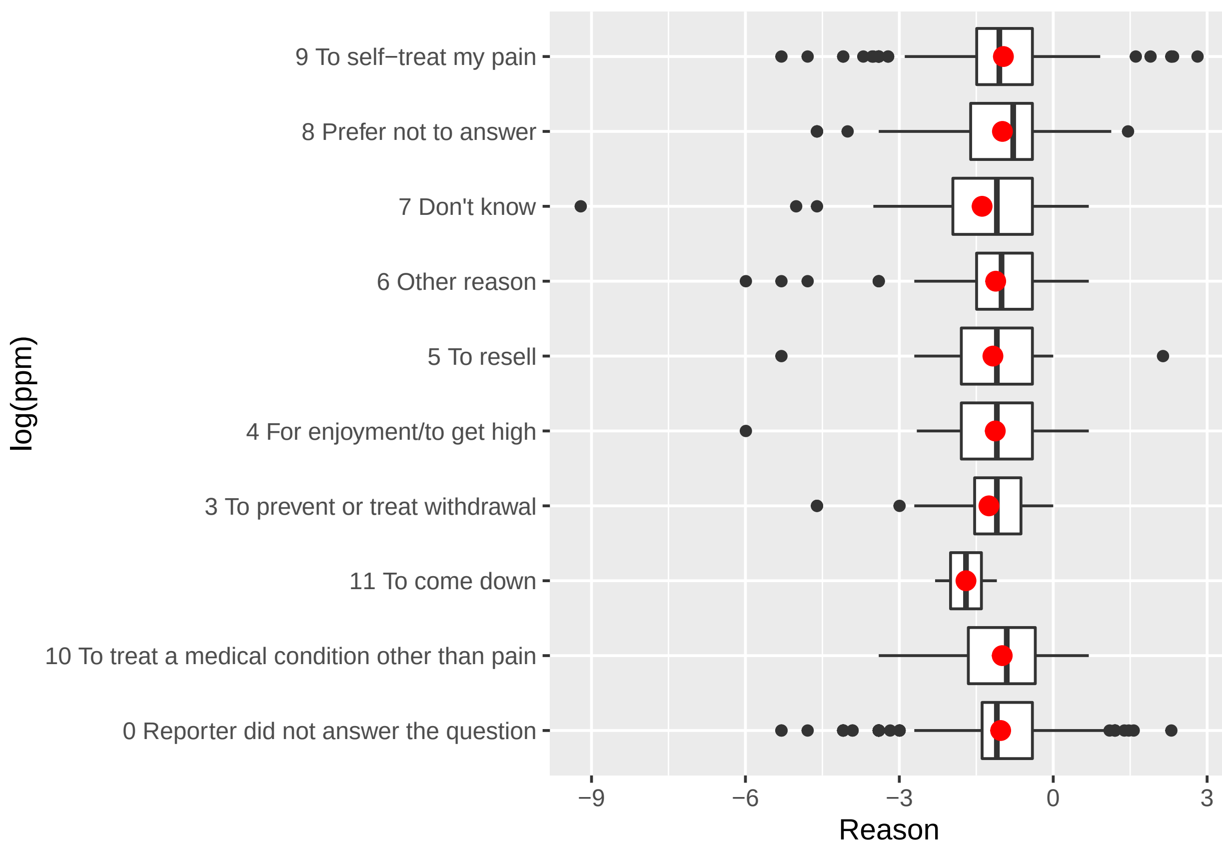
year & quarter vs.log(ppm)



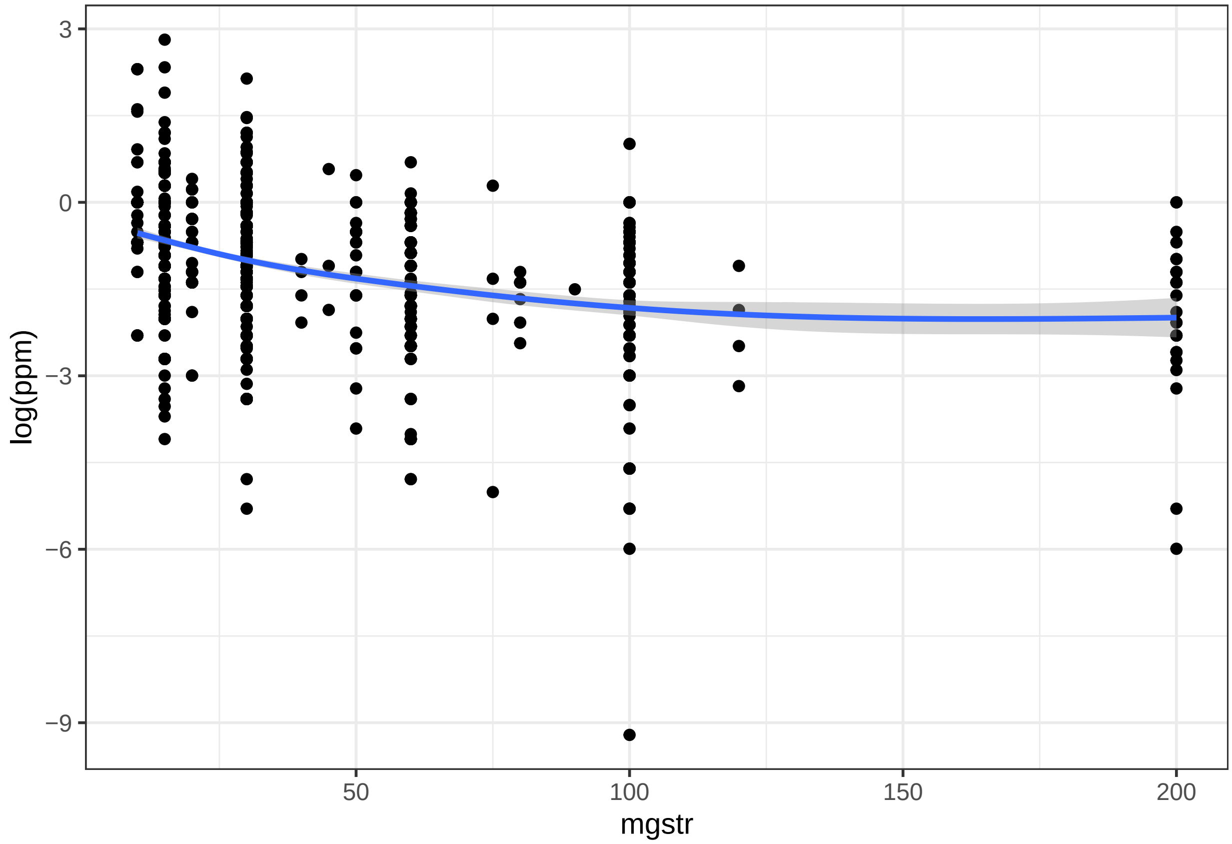
bulk_purchase vs.log(ppm)

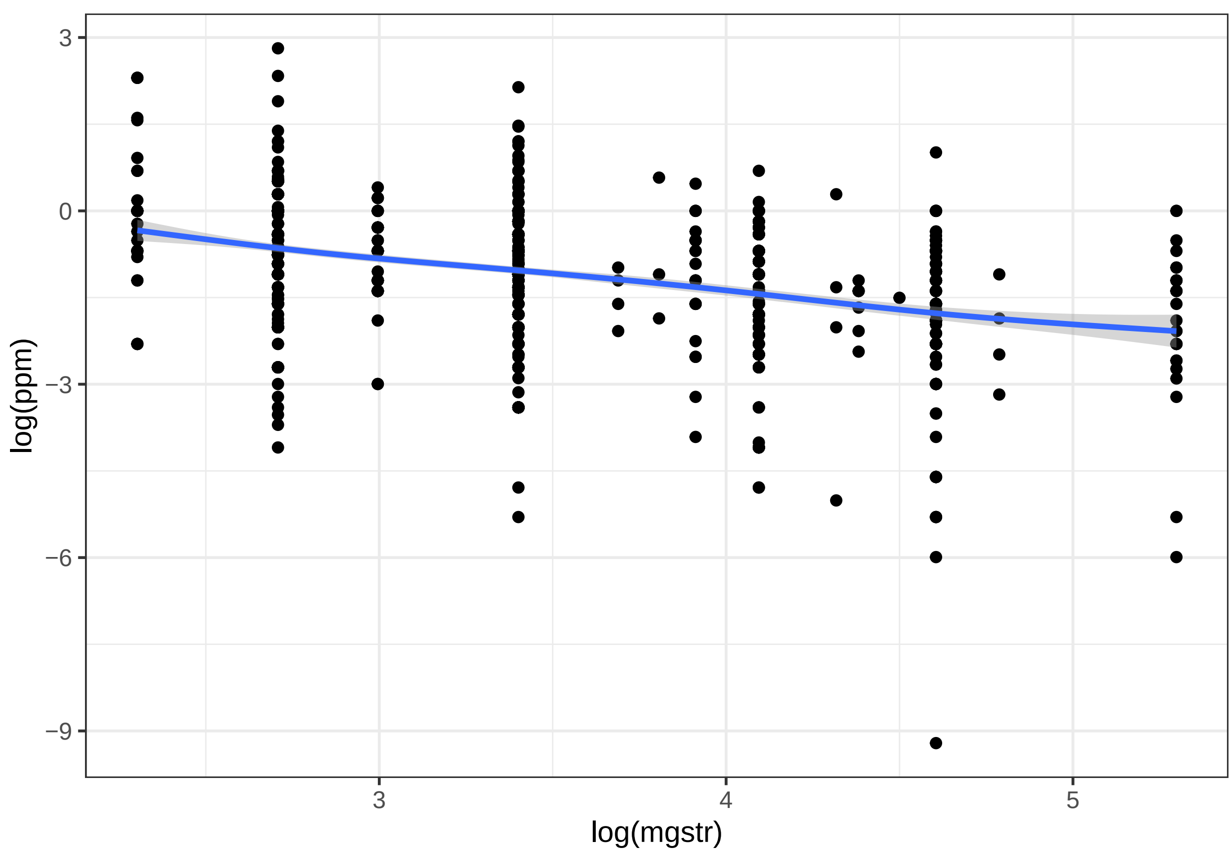


Primary_Reason vs.log(ppm)



mgstr vs. log(ppm)





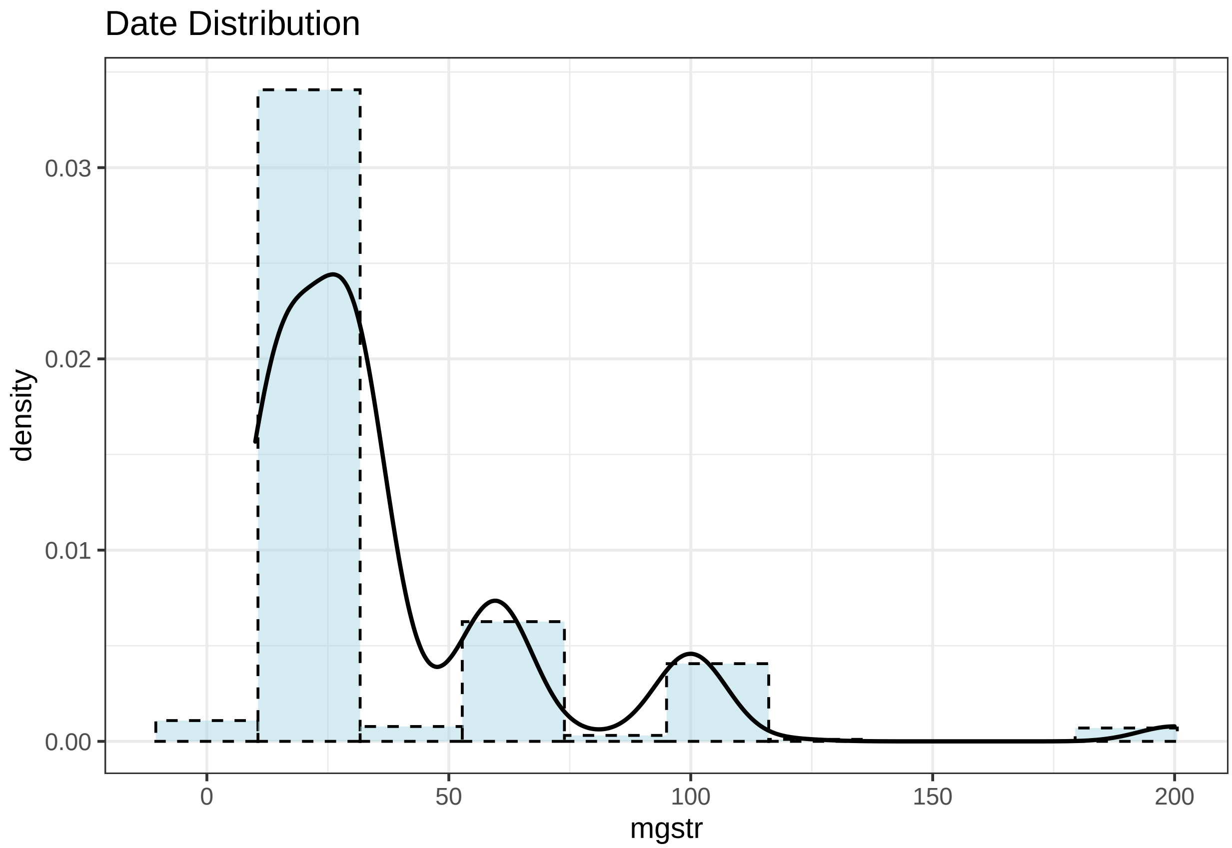
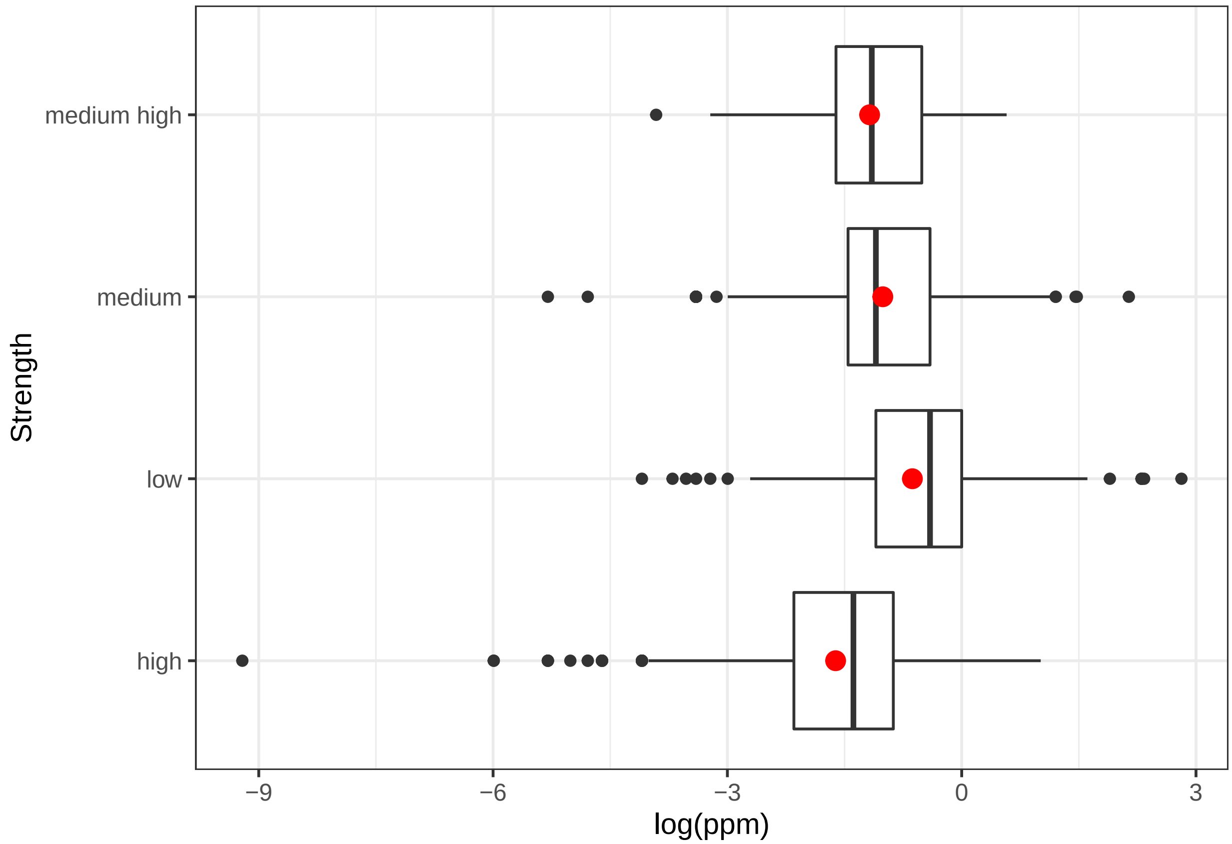


Table 3: Sample Size for mgstr Levels

10	15	20	30	40	45	50	60	75	80	90	100	120	200
42	572	47	698	4	3	23	242	4	7	1	157	4	27

	mgstr
25%	15
50%	30
75%	50



Model

Grouping	BIC
Basic	5231.237
\+ Source	5258.679
\+ Reason	5033.391
\+ Bulk	5041.267
\+ mgstr	5113.820
\+ quarter	5148.989

From this it looks like the best model includes date_diff, quarter, and mgstr

choose grouping variable

Grouping	BIC
State	5070.793
City	5079.681
Region	5078.169

Choose `State` as our grouping variable

```
## [1] 5070.793 5062.219 4964.250 4955.328
```

```
## Backward reduced random-effect table:
```

```
##
##              Eliminated npar  logLik    AIC    LRT Df Pr(>Chisq)
## <none>                    24 -2441.0 4929.9
## (1 | state)          0    23 -2445.6 4937.2 9.2376  1   0.002371 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Backward reduced fixed-effect table:
## Degrees of freedom method: Satterthwaite
##
##                 Eliminated  Sum Sq Mean Sq NumDF  DenDF F value  Pr(>F)
## source                   1   1.582   0.396     4 1816.6  0.4751 0.75403
## date_diff                2   0.115   0.115     1 1823.4  0.1385 0.70985
## primary_reason           3  12.517   1.391     9 1823.1  1.6689 0.09128 .
## quarter                  4   6.223   2.074     3 1828.0  2.4674 0.06047 .
## mgstr2                   0 246.985  82.328     3 1826.3 97.5703 < 2e-16 ***
## bulk_purchase            0   3.357   3.357     1 1827.4  3.9784 0.04624 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Model found:
## log(ppm) ~ (1 | state) + mgstr2 + bulk_purchase
```

```
## [1] 1.252409e-06
```

```
## [1] 1
```

```
## [1] 1.158104e-05
```

```
## [1] 1
```

```
## [1] 2.677778e-55
```

```
## [1] 1
```

We now have `quarter`, `bulk_purchase`, `primary_reason` and `mgstr2` in our model, regarding `state` as the grouping variable.

Why/how did we choose the variables to put into the model? too many predictors?

Unique values: state = 45 quarter = 4 bulk purchase = 2 primary reason = 10 mgstr = 14

only have 1831 observations

final model

```
## [1] 5036.09
```

```
## [1] 5026.693
```

log(ppm)

Predictors

Estimates

CI

p

(Intercept)

-1.63

-1.76 – -1.49

<0.001

mgstr2 [low]

0.98

0.86 – 1.09

<0.001

mgstr2 [medium]

0.60

0.49 – 0.71

<0.001

mgstr2 [medium high]

0.40

0.06 – 0.74

0.023

quarter [2]

0.07

-0.04 – 0.19

0.228

quarter [3]

0.16

0.04 – 0.27

0.010

quarter [4]

0.13

0.02 – 0.25

0.025

bulk_purchase [1 Bulkpurchase]

-0.10

-0.20 – -0.01

0.038

primary_reason [10 Totreat a medical conditionother than pain]

0.09

-0.22 – 0.40

0.578

primary_reason [11 Tocome down]

-0.64

-1.91 – 0.63

0.325

primary_reason [3 Toprevent or treatwithdrawal]

-0.26

-0.54 – 0.03

0.075

primary_reason [4 Forenjoyment/to get high]

-0.11

-0.35 – 0.12

0.354

primary_reason [5 Toresell]

-0.18

-0.48 – 0.12

0.246

primary_reason [6 Otherreason]

-0.03

-0.23 – 0.17

0.751

primary_reason [7 Don’tknow]

-0.29

-0.54 – -0.04

0.026

primary_reason [8 Prefernot to answer]

0.05

-0.07 – 0.18

0.376

primary_reason [9 Toself-treat my pain]

0.05

-0.06 – 0.15

0.416

Random Effects

 2

0.83

 00 state

0.01

ICC

0.02

N state

45

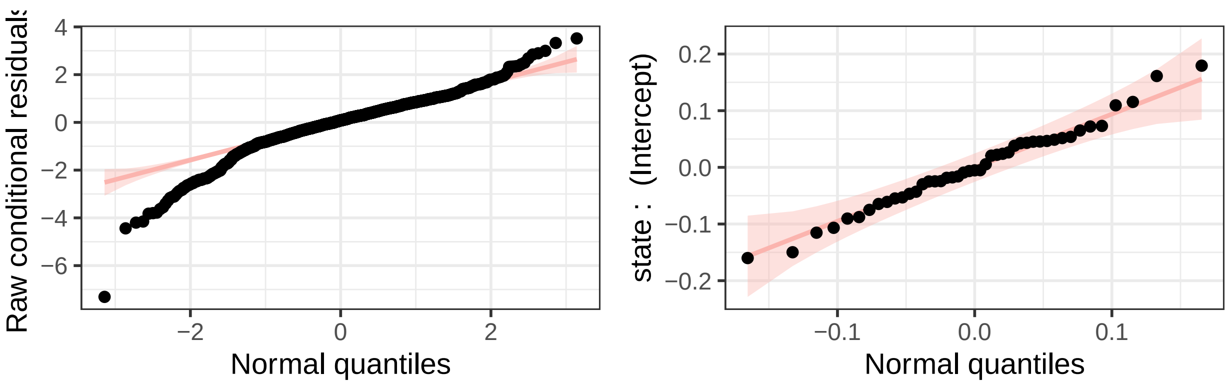
Observations

1831

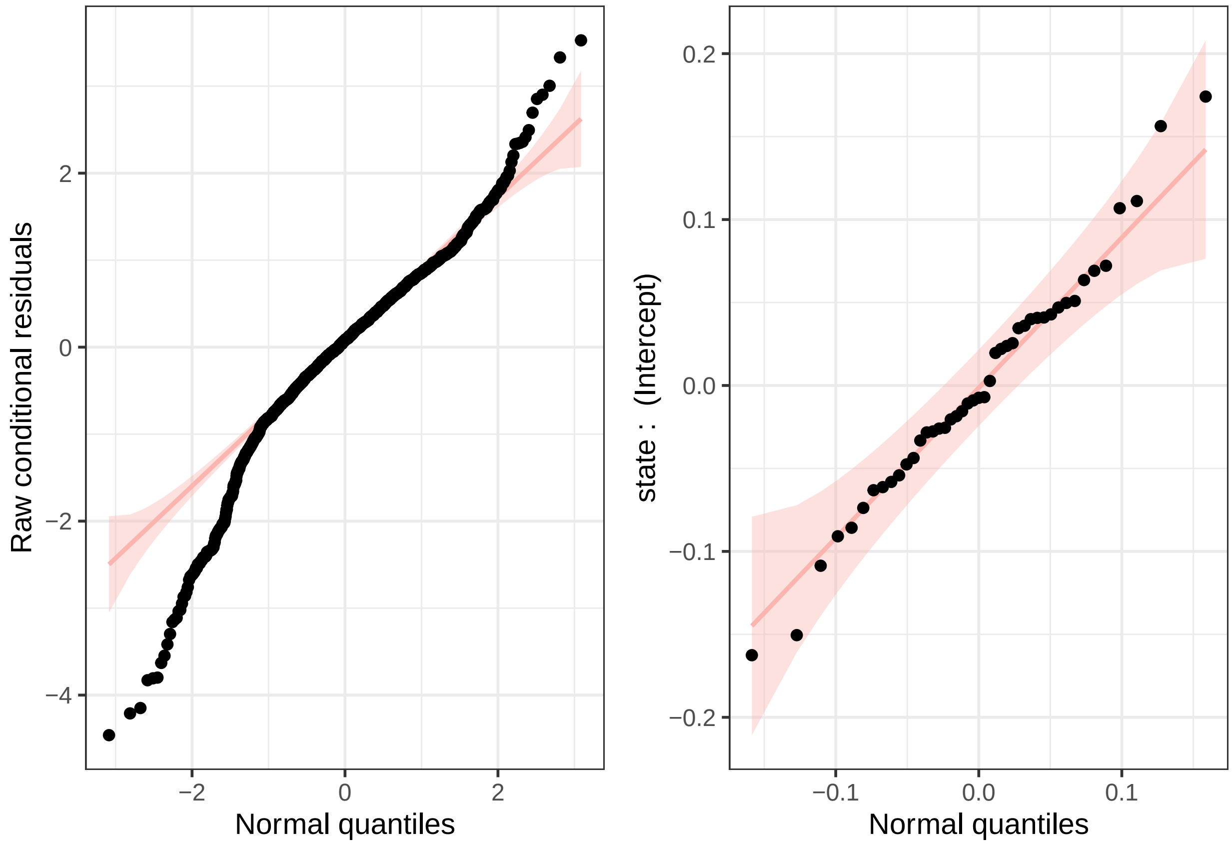
Marginal R2 / Conditional R2

0.149 / 0.164

```
## NULL
```



Remove the data point with the lowest residual.

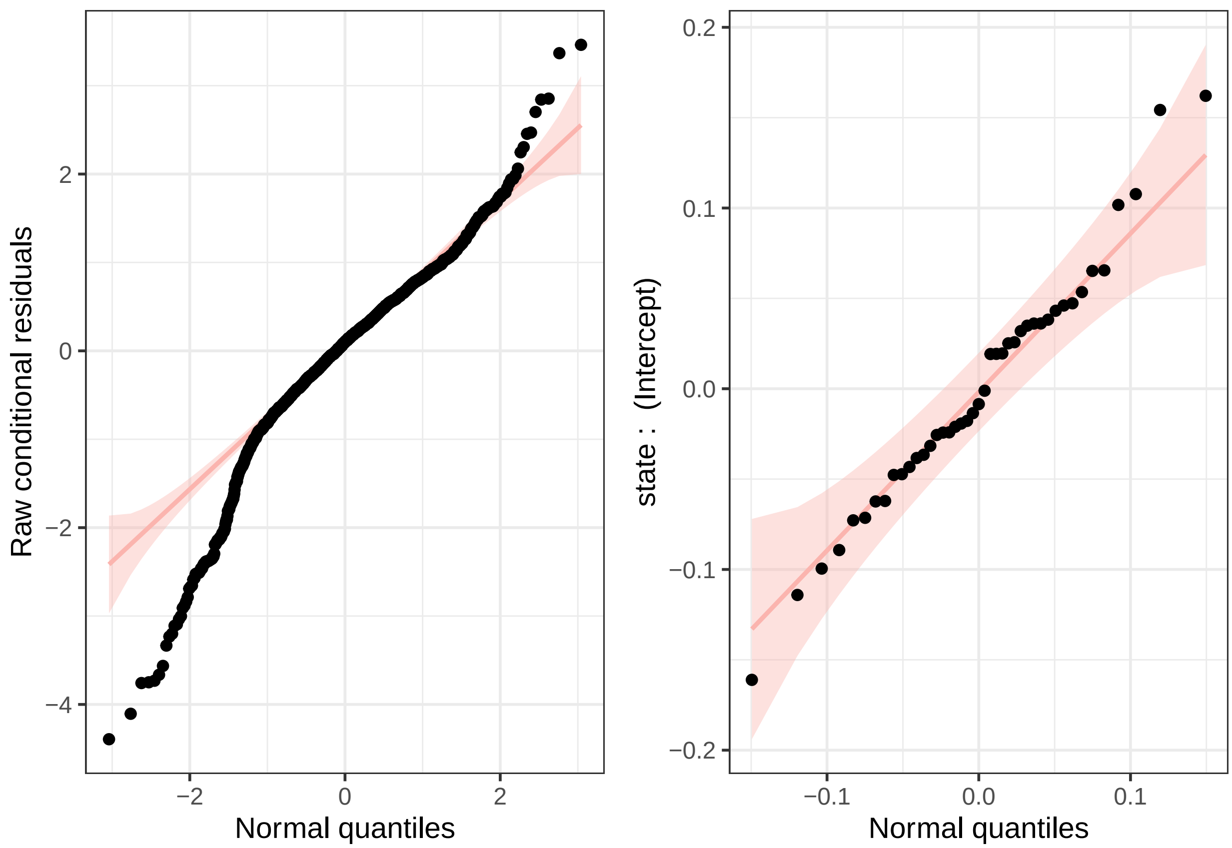


Influence

```
## integer(0)
```

```
## pdf
##   2
```

```
##    rownames.mod_final3_inf..fixed.effects..state... cooks_distance infindiv
## 1                                        California     0.13448366     TRUE
## 4                                           Georgia     0.09317747     TRUE
## 5                                           Florida     0.10427666     TRUE
## 13                                          Arizona     0.16956752     TRUE
## 27                                         Missouri     0.12189307     TRUE
```




```
## $state
```

```
## pdf
##   2
```

Plots are not good -> not remove those two states?

Interclass correlation is 0.0159, very small so very little correlation across states. Including bulk purchases, the interclass correlation is 0.016, so bulk purchase actually increases the heterogeneity across states by a very small amount.

Make table with results for all models tested in ANOVA