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.

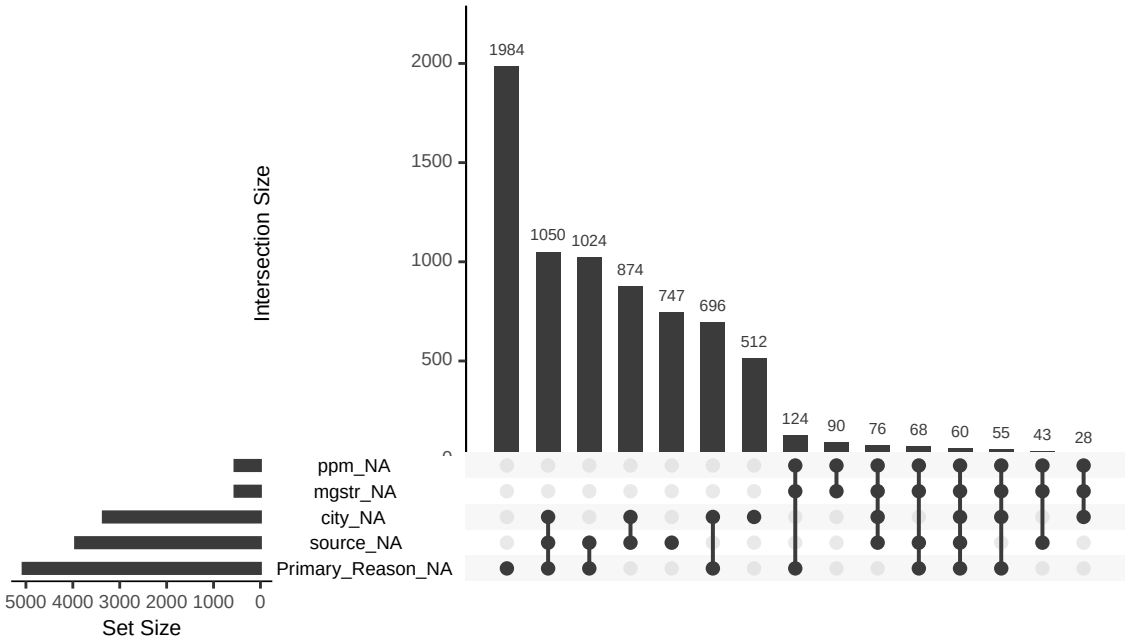
Jack - Programmer: Primarily responsible for all things coding. The programmer is responsible for putting everyone's code together and making sure the final product is "readable".

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

Introduction

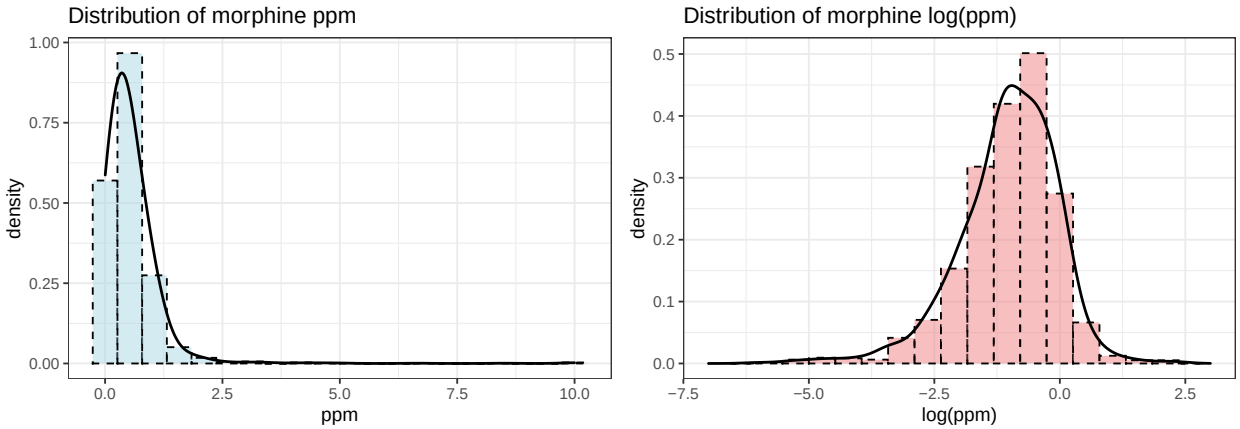
EDA

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)

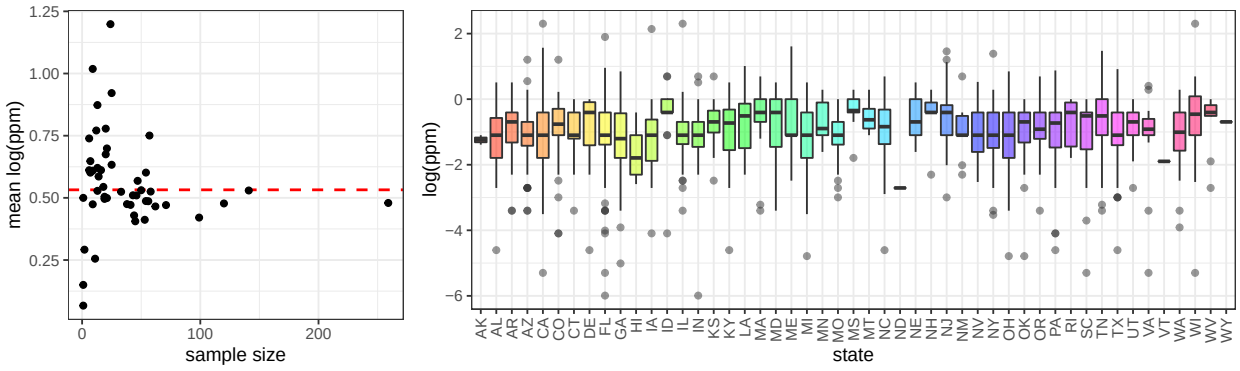
We see that there are 4 states that have a sample size of 1, North Dakota, Vermont, Washington DC, and Wyoming, as well as 1 state that has a sample size of 2, Alaska. Due to the extremely small sample sizes we decided to remove these states form our dataset to avoid computational instability.

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

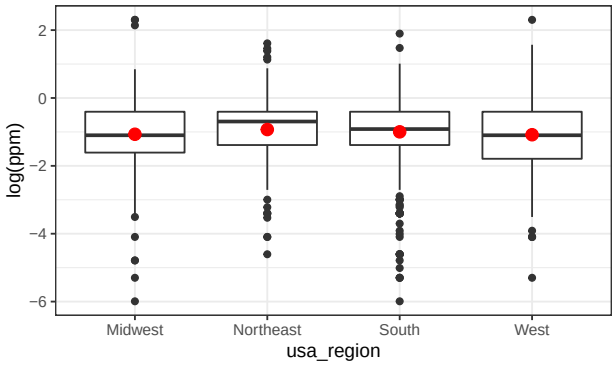


We observe that the within-state means for states with higher sample sizes in general adhere more closely to the grand mean. It is also evident that the log(ppm) distributions differ little as compared to the within-state variance. This is conducive to the borrowing of information between states.

region vs. log(ppm)

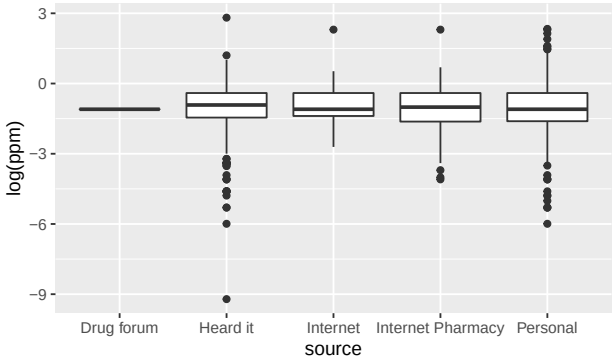
We also have access to the broader region in which a purchase is made. This could be useful if we wanted to develop a simpler model that still captured variation by purchase location.

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



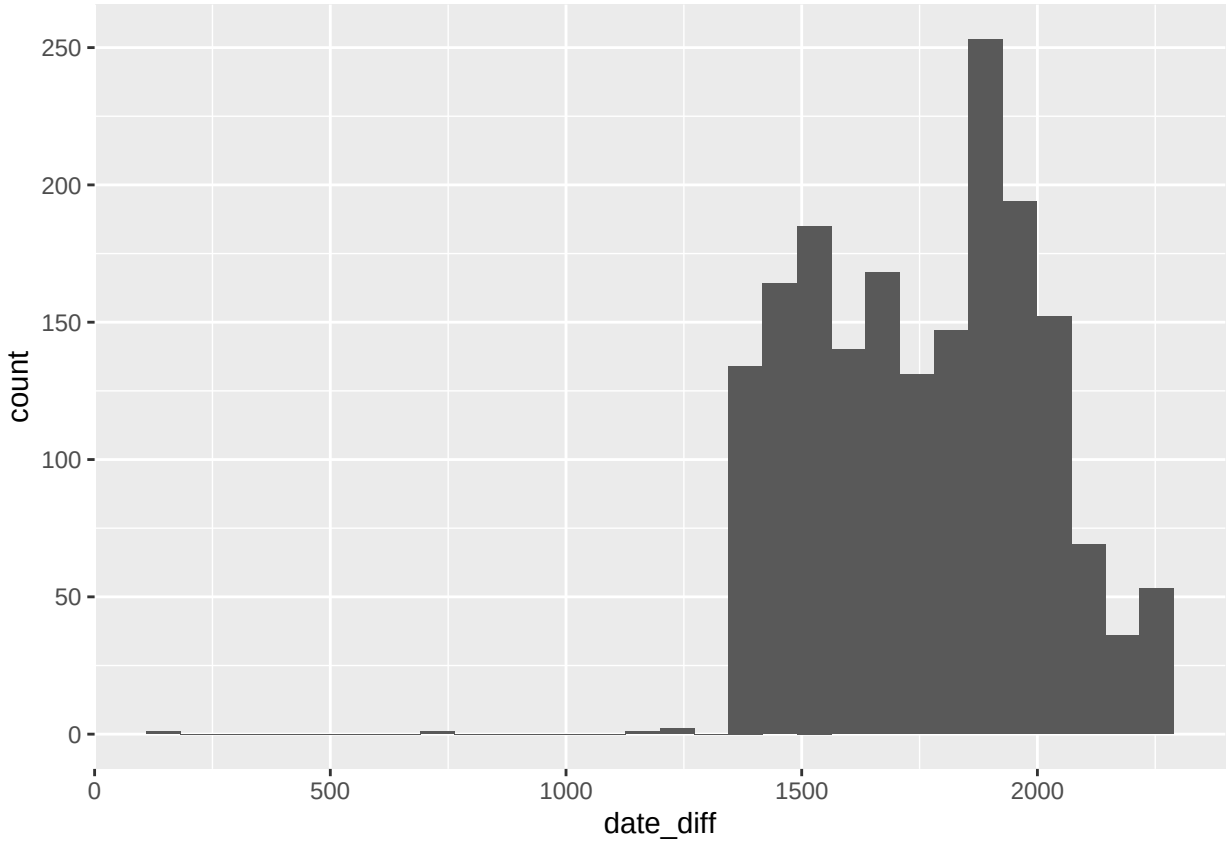
source vs. log(ppm)

	source	n	mean
1	Drug forum	1	−1.099
2	Heard it	578	−1.033
3	Internet	103	−0.985
4	Internet Pharmacy	48	−1.185
5	Personal	1101	−1.029



date

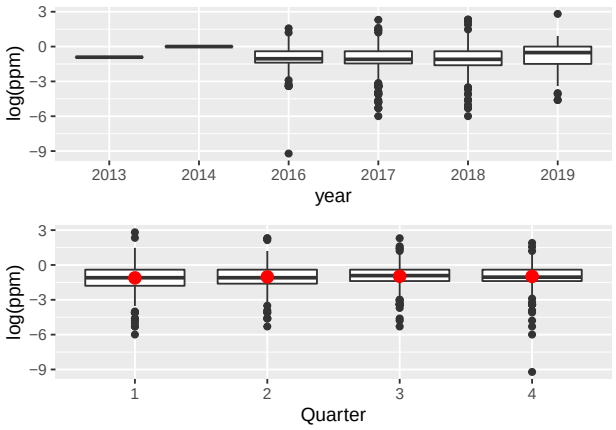
record `price_date` as a continuous variable counting days from some start date.



year & quarter vs.log(ppm)

	year	n
1	2013	1
2	2014	1
3	2016	233
4	2017	780
5	2018	748
6	2019	68

	quarter	n
1	1	575
2	2	430
3	3	391
4	4	435




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

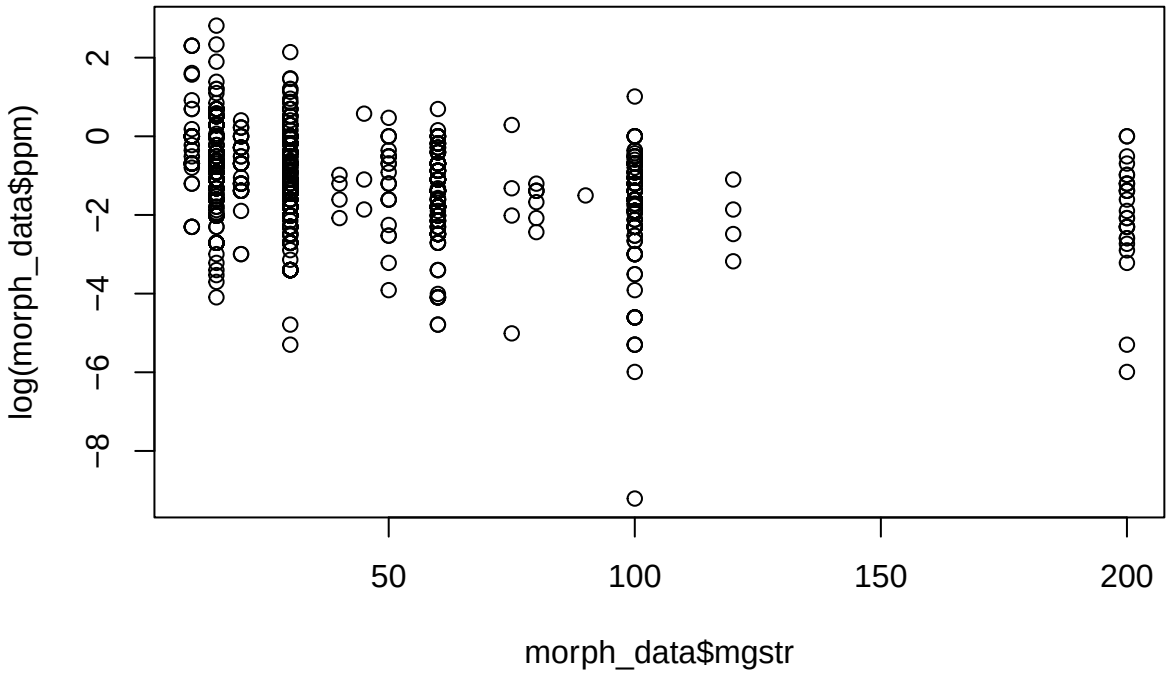
bulk_purchase vs.log(ppm)

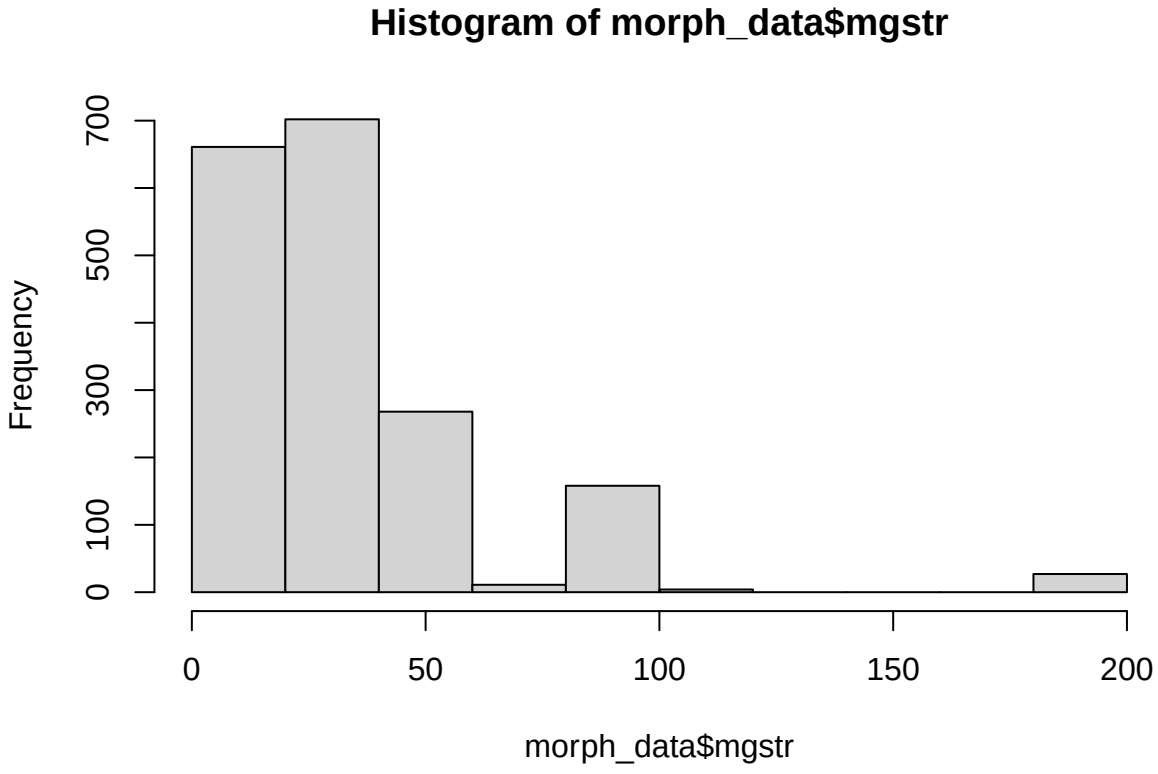
```
## pdf
##   2
```

Primary_Reason vs.log(ppm)

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

mgstr vs. log(ppm)





```
## # A tibble: 14 x 2
##    mgstr     n
##    <dbl> <int>
##  1    10    42
##  2    15   572
##  3    20    47
##  4    30   698
##  5    40     4
##  6    45     3
##  7    50    23
##  8    60   242
##  9    75     4
## 10    80     7
## 11    90     1
## 12   100   157
## 13   120     4
## 14   200    27
```

```
##   0%  25%  50%  75% 100%
##   10   15   30   50  200
```

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

Model

Grouping	BIC
All	5148.989
- Source	5113.820
- Reason	5041.267
- Bulk	5033.391
- mgstr	5258.679
- quarter	5231.237

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

choose grouping variable

```
##   0%  25%  50%  75% 100%
##   10   15   30   50  200
```

Grouping	BIC
State	5064.901
City	5073.967
Region	5072.387

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

Choose `State` as our grouping variable

```
## [1] 5064.901 5049.336 4958.986 4942.195
```

```
## Backward reduced random-effect table:
##
##             Eliminated npar  logLik    AIC    LRT Df Pr(>Chisq)
## <none>                   24 -2434.5 4917.1
## (1 | state)          0   23 -2439.2 4924.3 9.2676  1   0.002332 **
## ---
## 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   1.684   0.421     4 1814.2   0.5078 0.73004
## date_diff               2   0.370   0.370     1 1820.6   0.4462 0.50421
## primary_reason          3  12.217   1.357     9 1820.8   1.6353 0.09992 .
## quarter                 4   5.830   1.943     3 1825.8   2.3208 0.07347 .
## mgstr2                  0 253.916  84.639     3 1822.1 100.7387 < 2e-16 ***
## bulk_purchase           0   3.368   3.368     1 1825.1   4.0085 0.04542 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Model found:
## log(ppm) ~ (1 | state) + mgstr2 + bulk_purchase
```

```
## [1] 0.5346387
```

```
## [1] 0.04501256
```

```
## [1] 0.03605441
```

```
## [1] 0.08530521
```

```
## [1] 1.26354e-59
```

```
## [1] 0.7304275
```

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] 5030.586
```

```
## [1] 5014.254
```

log(ppm)

Predictors

Estimates

CI

p

(Intercept)

-1.60

-1.73 – -1.47

<0.001

mgstr2 [low]

1.41

1.12 – 1.70

<0.001

mgstr2 [medium]

0.91

0.80 – 1.02

<0.001

mgstr2 [medium high]

0.56

0.46 – 0.67

<0.001

quarter [2]

0.06

-0.06 – 0.17

0.345

quarter [3]

0.15

0.03 – 0.27

0.013

quarter [4]

0.13

0.01 – 0.24

0.033

bulk_purchase [1 Bulkpurchase]

-0.10

-0.20 – -0.01

0.037

primary_reason [10 Totreat a medical conditionother than pain]

0.09

-0.22 – 0.39

0.584

primary_reason [11 Tocome down]

-0.62

-1.90 – 0.65

0.337

primary_reason [3 Toprevent or treatwithdrawal]

-0.25

-0.53 – 0.03

0.083

primary_reason [4 Forenjoyment/to get high]

-0.10

-0.34 – 0.14

0.398

primary_reason [5 Toresell]

-0.16

-0.46 – 0.15

0.309

primary_reason [6 Otherreason]

-0.04

-0.24 – 0.16

0.671

primary_reason [7 Don'tknow]

-0.29

-0.54 – -0.04

0.025

primary_reason [8 Prefernot to answer]

0.05

-0.07 – 0.17

0.389

primary_reason [9 Toself-treat my pain]

0.05

-0.06 – 0.16

0.353

Random Effects

2

0.83

00 state

0.01

ICC

0.02

N state

45

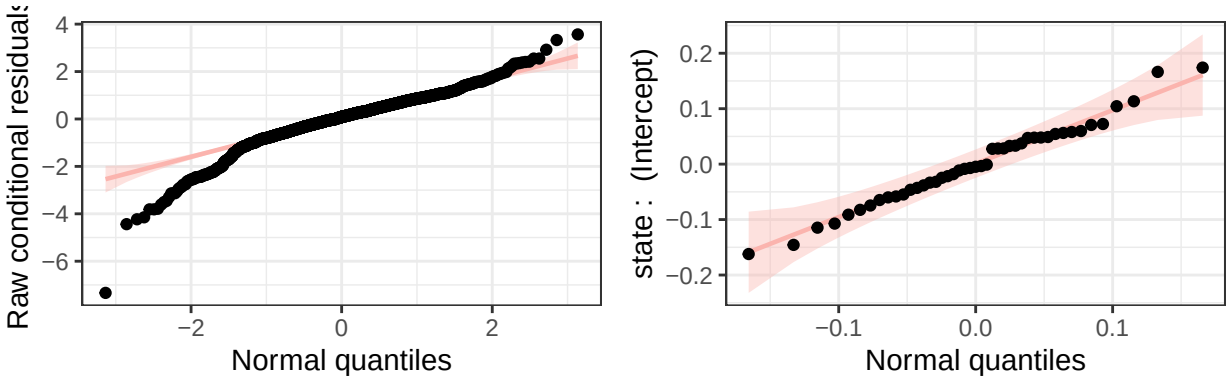
Observations

1829

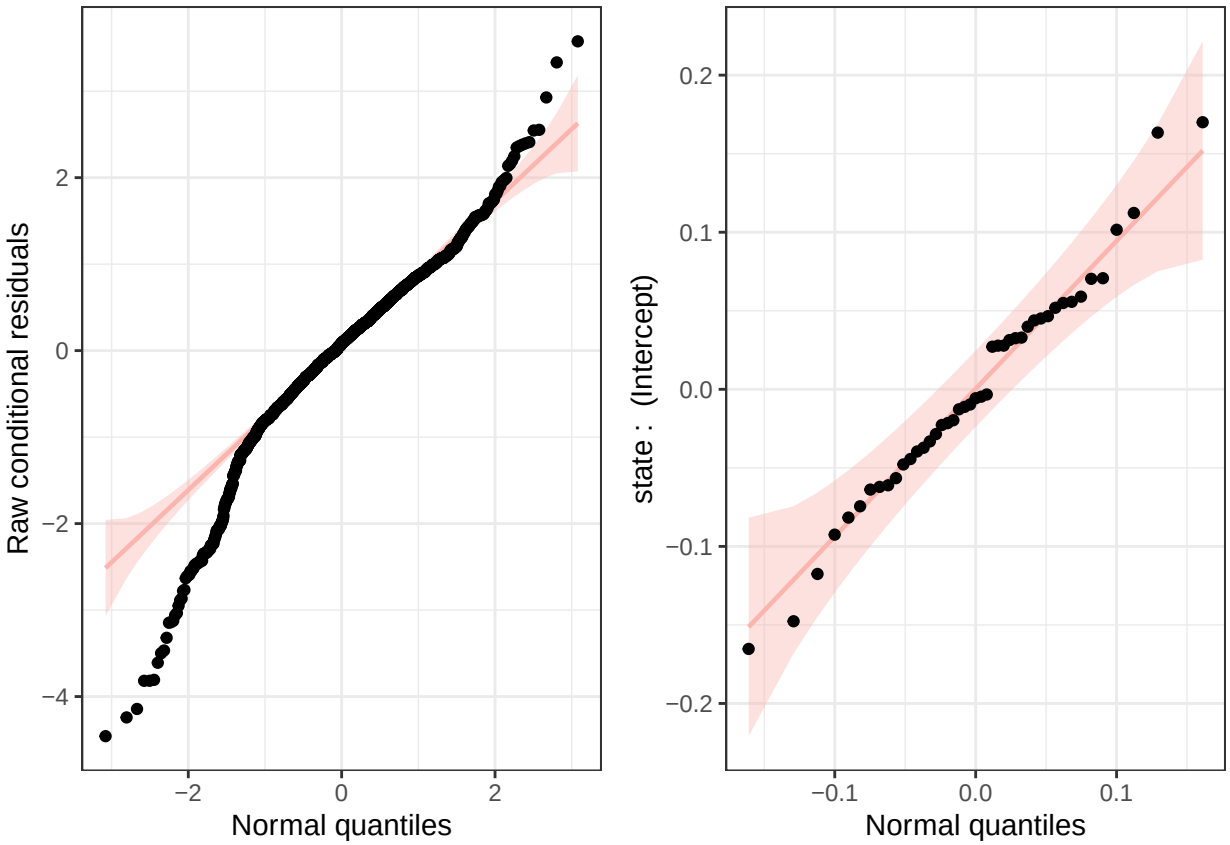
Marginal R2 / Conditional R2

0.152 / 0.167

```
## NULL
```



Remove the data point with the lowest residual.

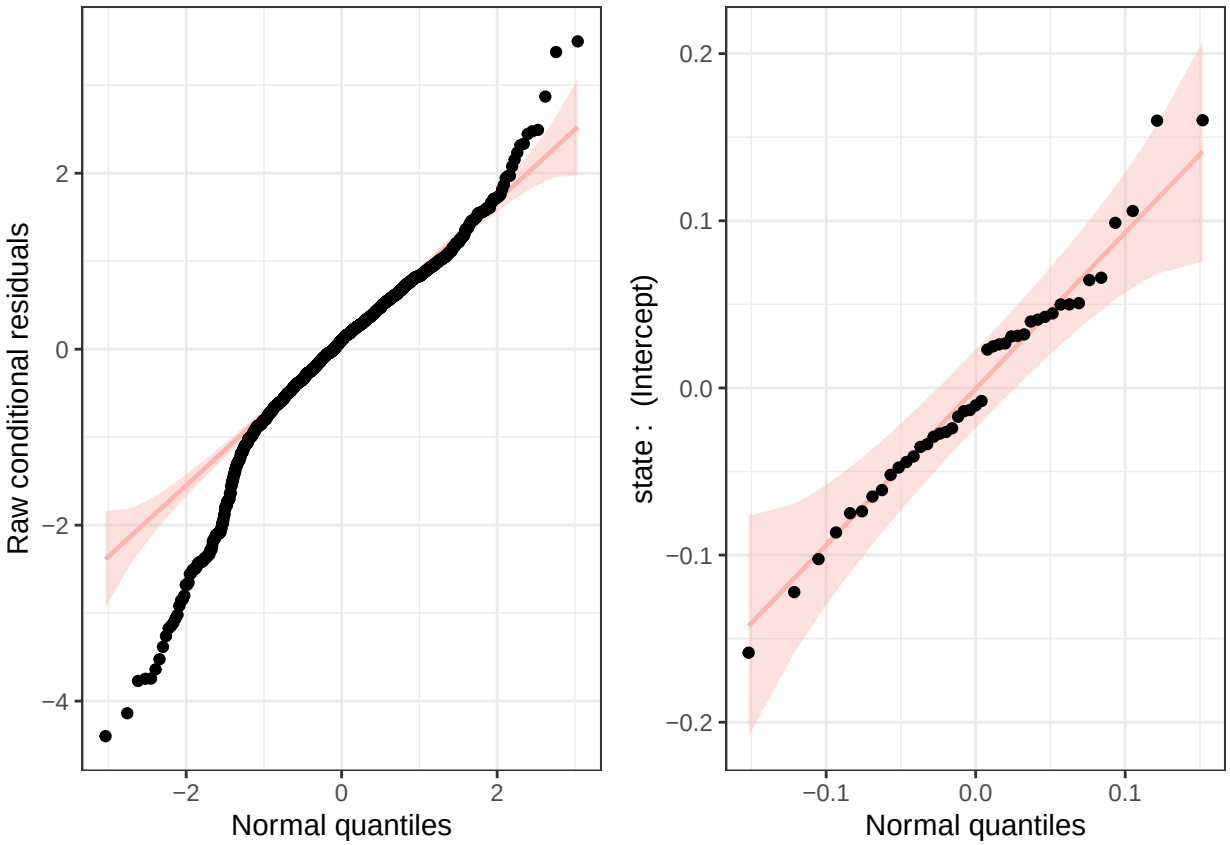


Influence

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

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

```
##    rownames.mod_final3_inf..fixed.effects..state... cooks_distance infindiv
## 1                                        California     0.14898010     TRUE
## 5                                           Florida     0.09397719     TRUE
## 13                                          Arizona     0.16303884     TRUE
## 27                                         Missouri     0.11458906     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