STA610 Case Study 1

Emily Gentles, Weiyi Liu, Jack McCarthy, Qinzhe Wang

06 October, 2021

Qinzhe - 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.

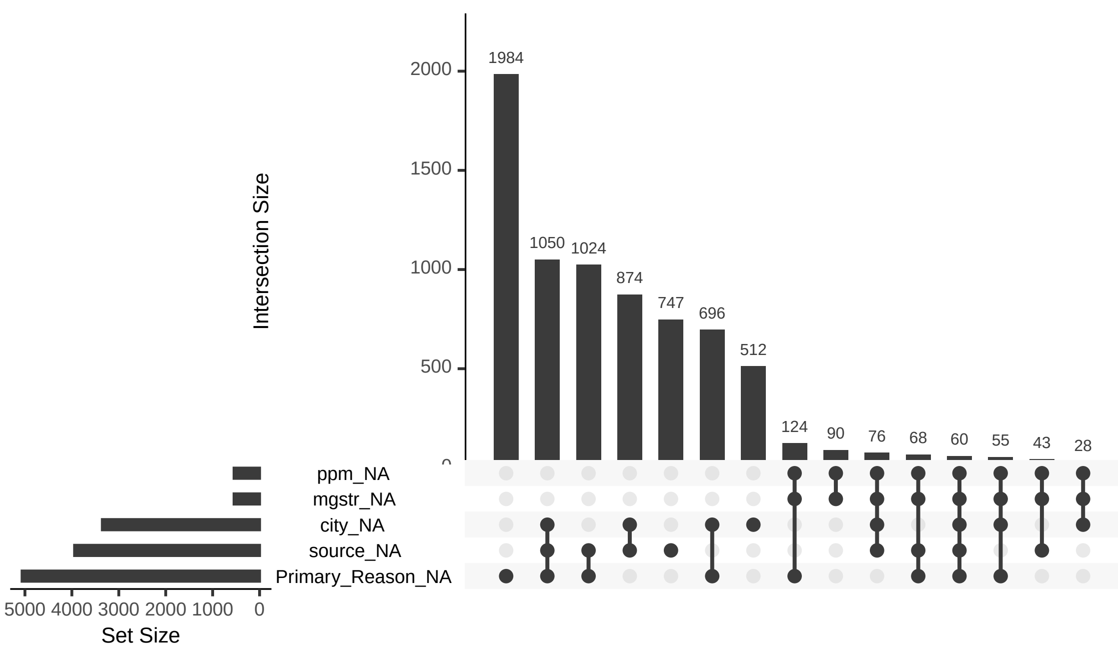
Jack - Programmer: Primarily responsible for all things coding. The programmer is responsible for putting everyone's code together and making sure the final product is "readable".

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

Introduction

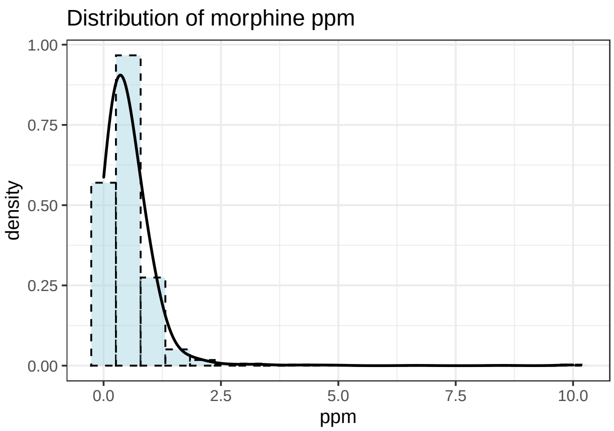
EDA

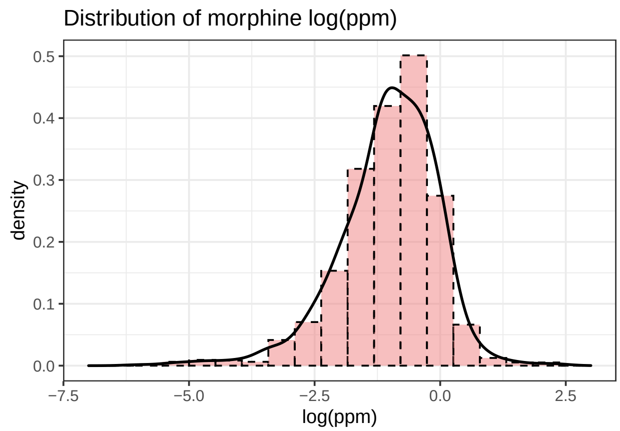
Missing Values



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)

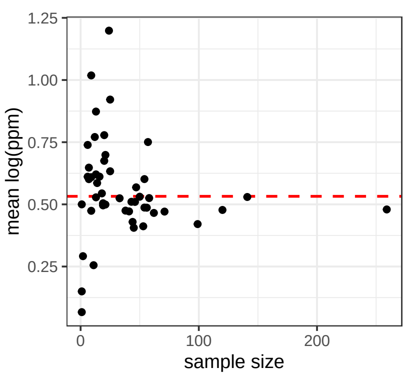
We see that there are 4 states that have a sample size of 1, North Dakota, Vermont, Washington DC, and Wyoming, as well as 1 state that has a sample size of 2, Alaska. Due to the extremely small sample sizes we decided to remove these states form our dataset to avoid computational instability.

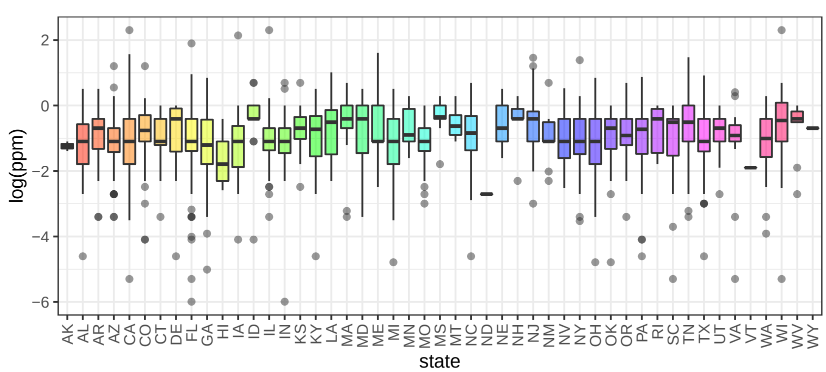
Table 1: 7 states with smallest sample size

North Dakota	Vermont	Washington, DC	Wyoming	Alaska
1	1	1	1	2

Table 2: 7 states with largest sample size

Arizona	Michigan	Texas	Florida	California
71	99	120	141	259



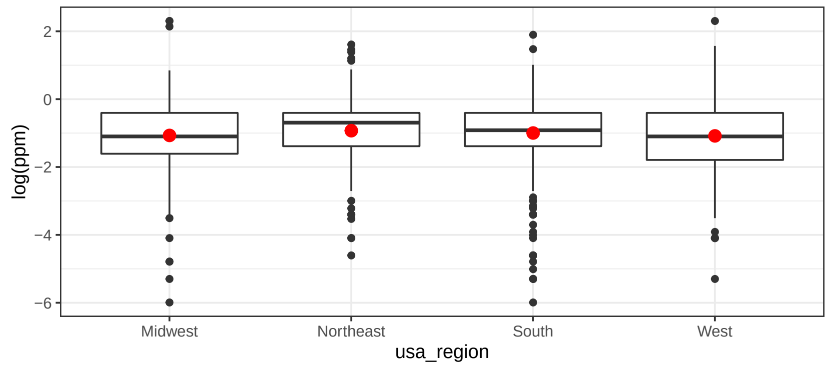


We observe that the within-state means for states with higher sample sizes in general adhere more closely to the grand mean. It is also evident that the log(ppm) distributions differ little as compared to the within-state variance. This is conducive to the borrowing of information between states.

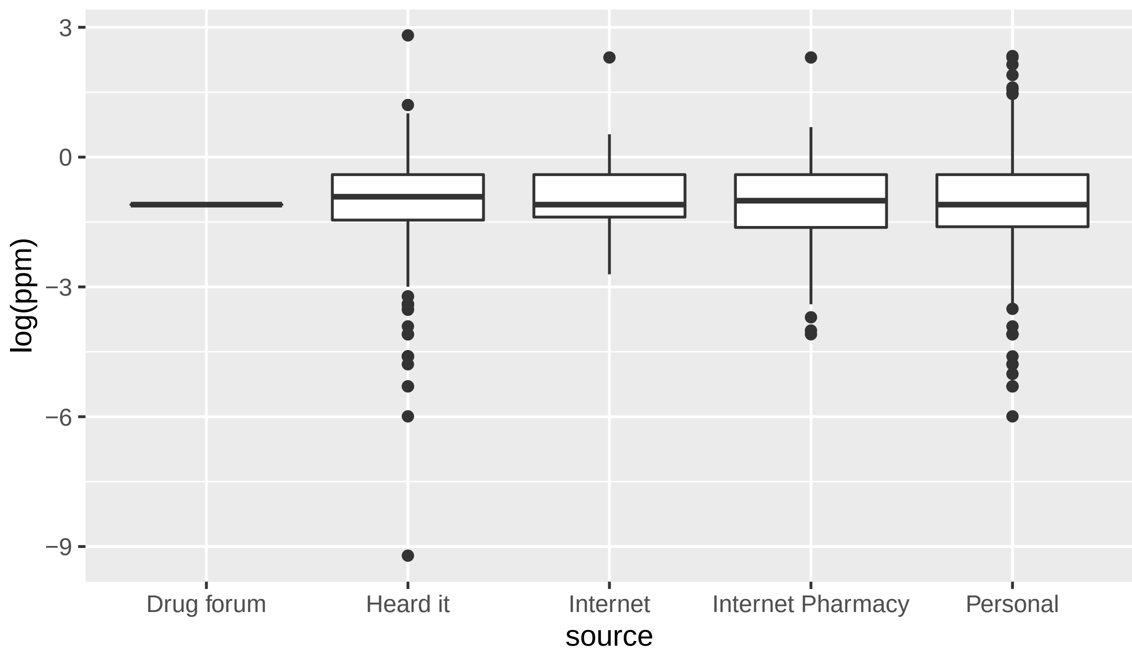
region vs. log(ppm)

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.

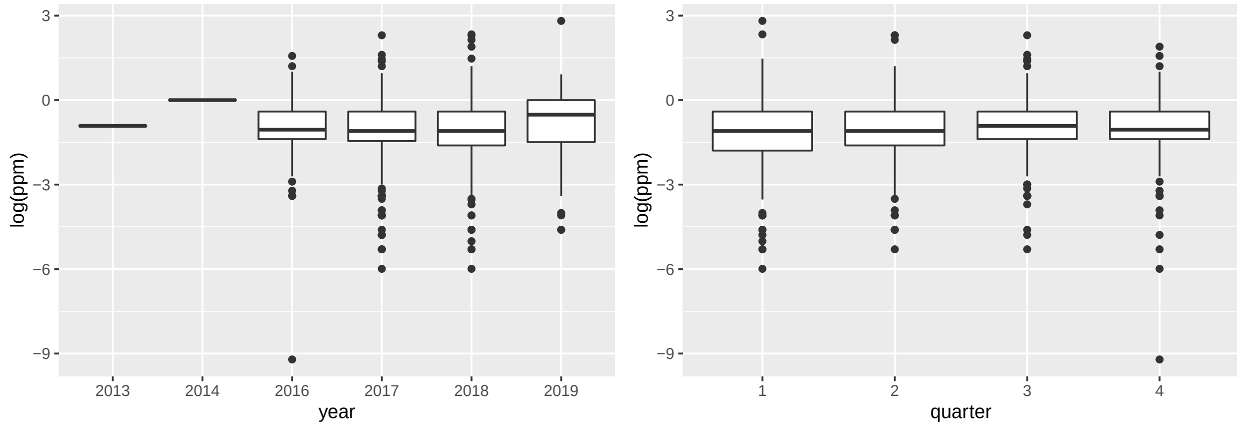
	usa_region	n	mean
1	Midwest	386	−1.069
2	Northeast	191	−0.930
3	South	673	−0.998
4	West	583	−1.083



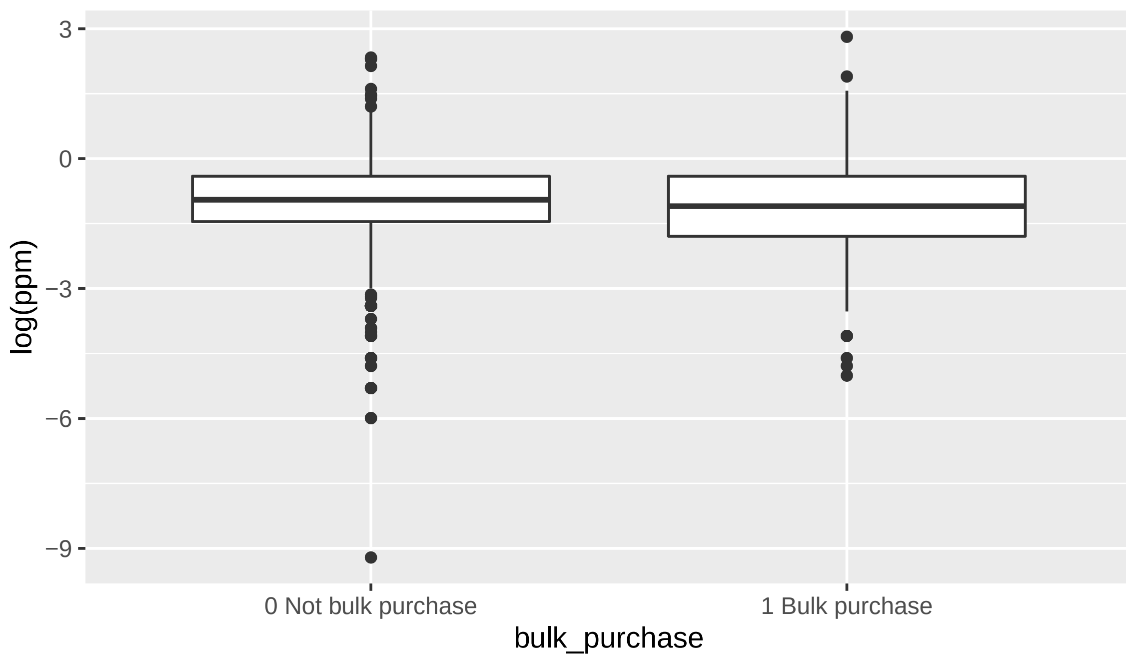
source vs. log(ppm)



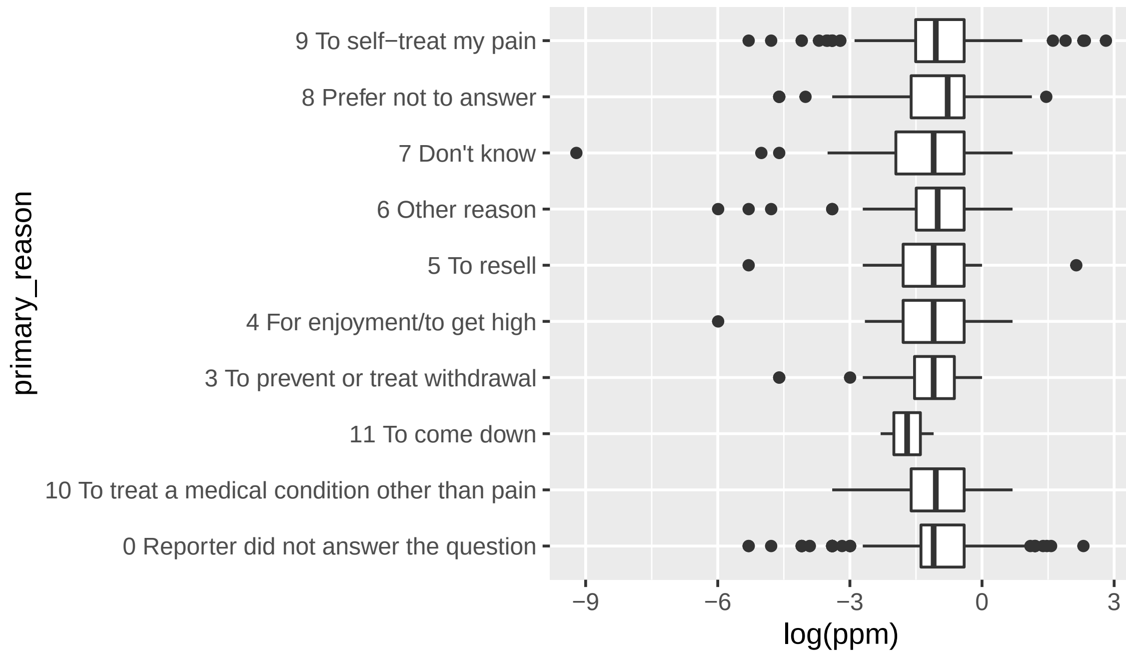
year & quarter vs.log(ppm)



bulk_purchase vs.log(ppm)



Primary__Reason vs.log(ppm)



Model

sth. wrong

```
## Data: morph_data
## Models:
## model1: log(ppm) ~ (1 | state)
## model2: log(ppm) ~ bulk_purchase + (1 | state)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model1    3 5201.8 5218.3 -2597.9   5195.8
## model2    4 5200.8 5222.8 -2596.4   5192.8 3.0152  1    0.08249 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Data: morph_data
## Models:
## model2: log(ppm) ~ bulk_purchase + (1 | state)
## model3: log(ppm) ~ -1 + bulk_purchase + (1 | state)
##        npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## model2    4 5200.8 5222.8 -2596.4   5192.8
## model3    4 5200.8 5222.8 -2596.4   5192.8     0  0

## Data: morph_data
## Models:
## model2: log(ppm) ~ bulk_purchase + (1 | state)
## model4: log(ppm) ~ bulk_purchase + source + (1 | state)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2    4 5200.8 5222.8 -2596.4   5192.8
```

```
## model4    8 5207.6 5251.7 -2595.8   5191.6 1.1984  4     0.8784
```

```
## Data: morph_data
## Models:
## model2: log(ppm) ~ bulk_purchase + (1 | state)
## model5: log(ppm) ~ bulk_purchase + source + year + (1 | state)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2    4 5200.8 5222.8 -2596.4   5192.8
## model5   13 5213.9 5285.7 -2594.0   5187.9 4.8328  9     0.8486
```

```
## Data: morph_data
## Models:
## model2: log(ppm) ~ bulk_purchase + (1 | state)
## model6: log(ppm) ~ bulk_purchase + source + quarter + (1 | state)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2    4 5200.8 5222.8 -2596.4   5192.8
## model6   11 5207.9 5268.5 -2592.9   5185.9 6.9289  7     0.4363
```

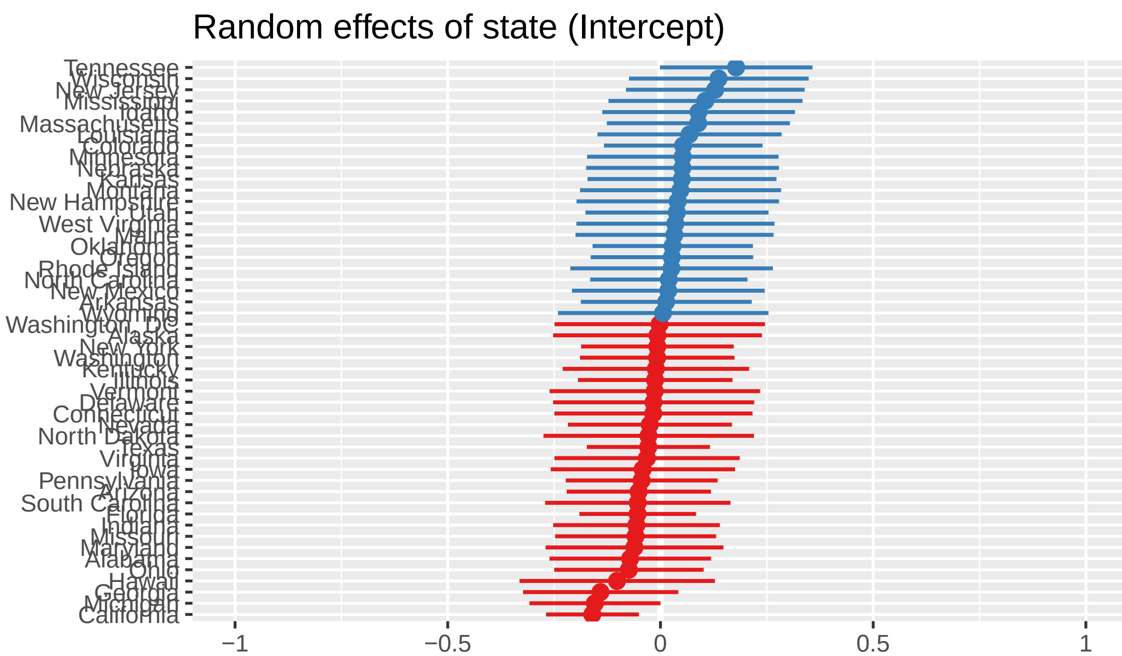
```
## Data: morph_data
## Models:
## model2: log(ppm) ~ bulk_purchase + (1 | state)
## model7: log(ppm) ~ bulk_purchase + source + (1 | year) + (1 | state)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2    4 5200.8 5222.8 -2596.4   5192.8
## model7    9 5209.6 5259.2 -2595.8   5191.6 1.1984  5      0.945
```

final model

log(ppm) ~ bulk_purchase + (1 | state)

consider using BIC

continue. . .




```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(ppm) ~ -1 + bulk_purchase + (1 | state)
##    Data: morph_data
##
## REML criterion at convergence: 5201.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -8.2634 -0.4819  0.0343  0.6322  3.9277
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  state    (Intercept) 0.01613  0.1270
##  Residual             0.97910  0.9895
## Number of obs: 1837, groups:  state, 50
##
## Fixed effects:
##                                   Estimate Std. Error t value
## bulk_purchase0 Not bulk purchase -0.97237    0.03526  -27.57
## bulk_purchase1 Bulk purchase     -1.06556    0.05187  -20.55
##
## Correlation of Fixed Effects:
##             b_0Nbp
## blk_prch1Bp 0.289

##                                    Estimate Std. Error   t value
## bulk_purchase0 Not bulk purchase -0.9723734 0.03526222 -27.57550
## bulk_purchase1 Bulk purchase     -1.0655603 0.05186505 -20.54486

##                                        2.5 %     97.5 %
```

```
## .sig01                           0.05007826  0.2043788
## .sigma                           0.95777569  1.0224962
## bulk_purchase0 Not bulk purchase -1.04215868 -0.8978858
## bulk_purchase1 Bulk purchase     -1.16831675 -0.9619632
```

Influence

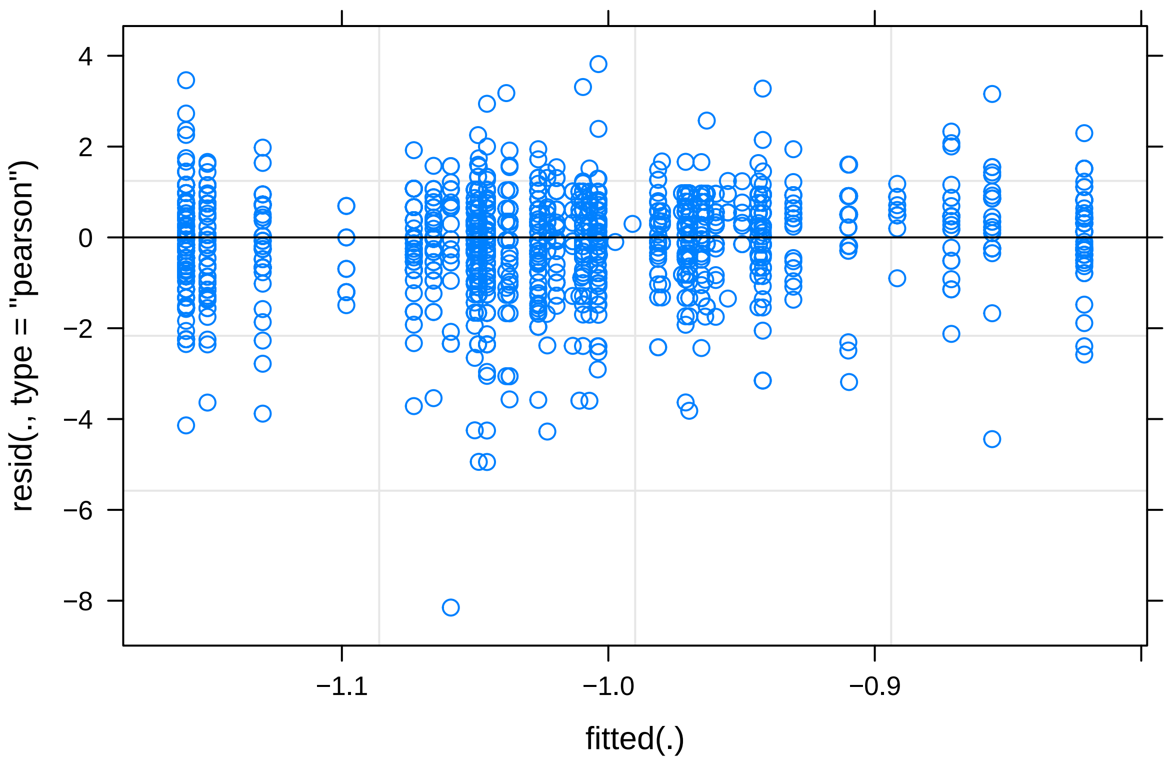
Heterogeneity across states

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

```
## [1] 0.98028984 0.01593162 0.01599205
```

```
## [1] 0.97910446 0.01613374 0.01621094
```



Model diagnostics