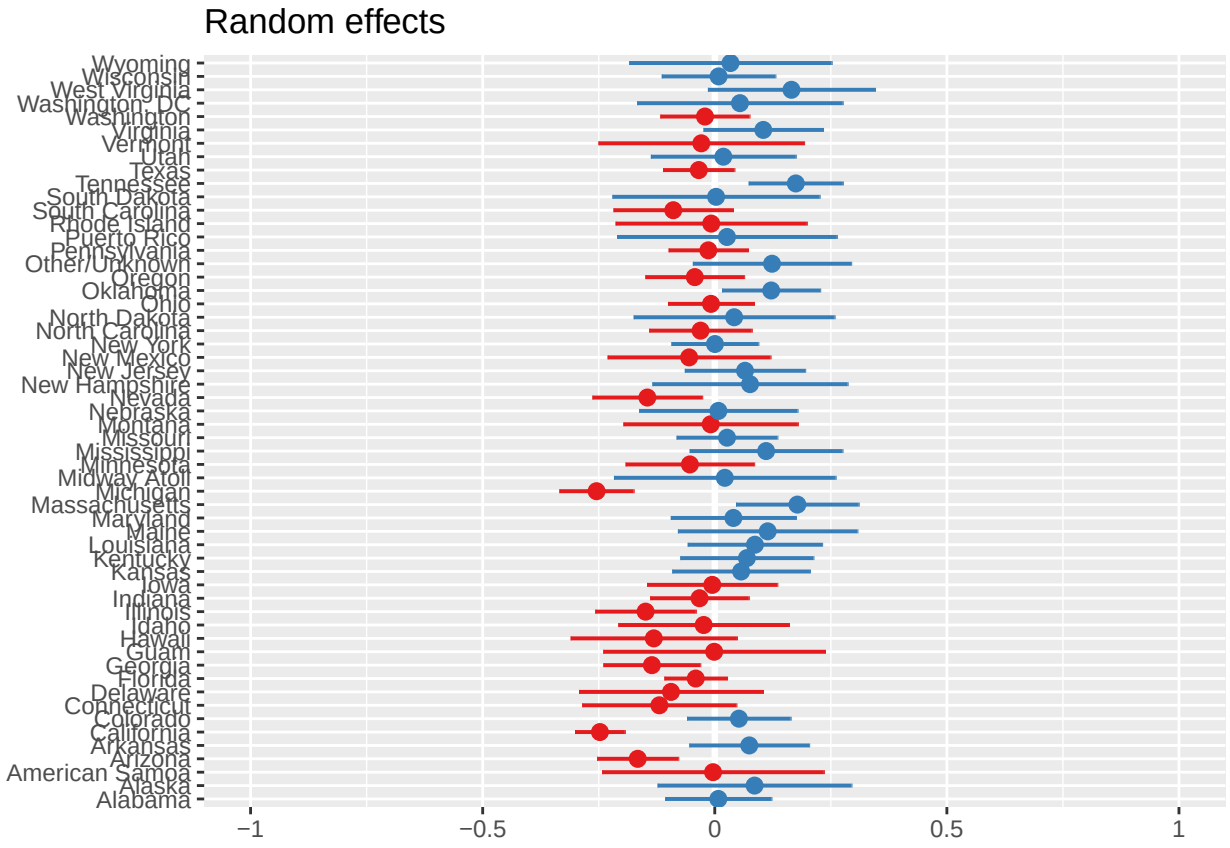
Model 1&2

Conclusion

Table 1: Fixed effect

	coef	Low CI	Up CI
Intecept	-0.590	-0.6390000	-0.5410000
mgstr	-0.010	-0.0105943	-0.0094057
Bulk	-0.103	-0.1539600	-0.0520400

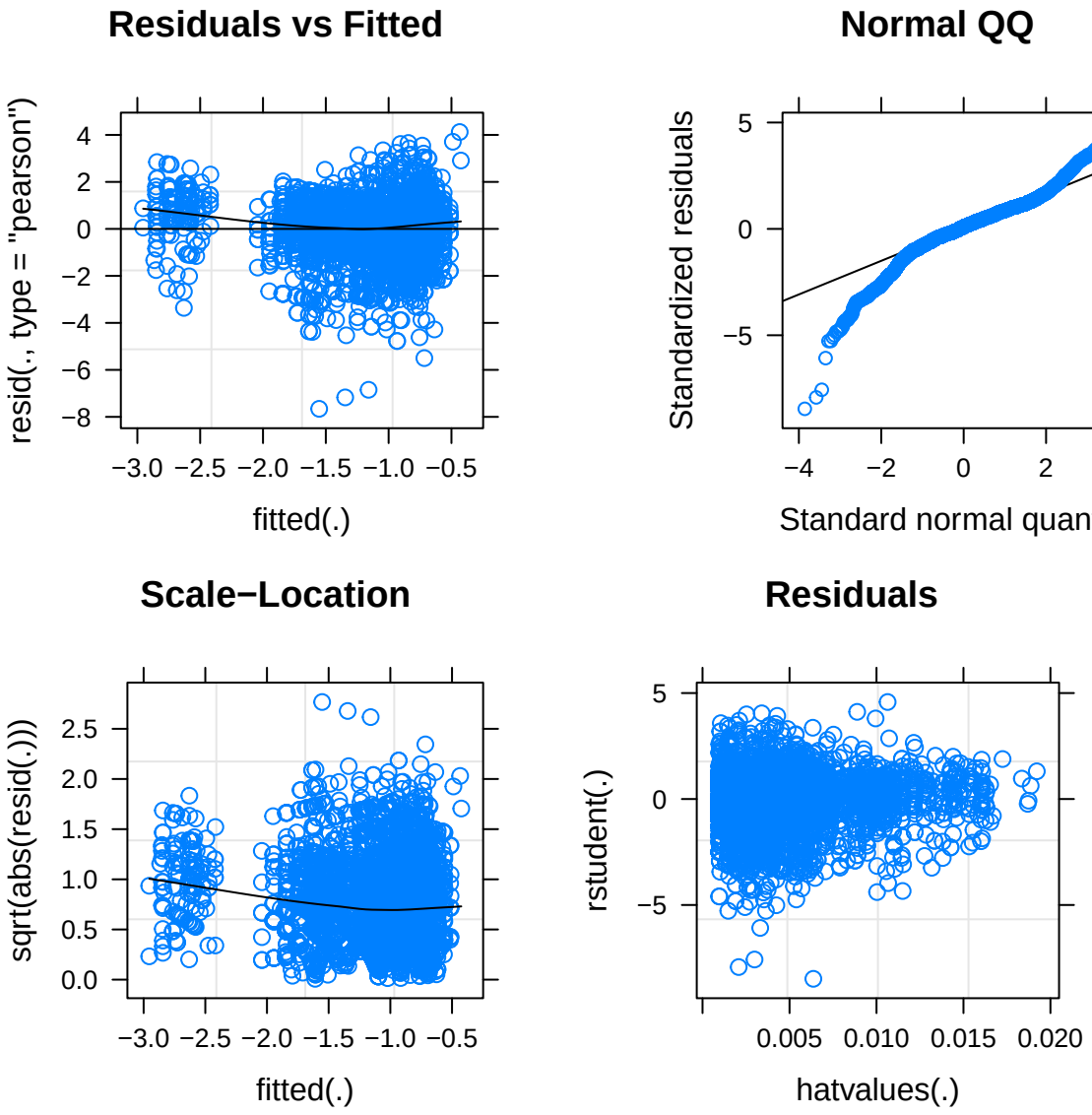


Our model only includes three variables: State, mgstr, and Bulk. Mgstr and Bulk are fixed effects and State has a random effect.

Let's first check the fixed effect in this model. As we can see form the fixed effect table, the intercept for this model is -0.59, which means the price of the drug would be $e^{-0.59} = 0.554$ while the dosage strength is 0 and Bulk is also 0. For mgstr, its coefficent is -0.01, which means a 1 unit increase in dosage strength will lead to a 0.01 decrease in log price per mg while holding all other coefficients constant. That is to say, a 1 unit increase in mgstr will lead to original price increase by a factor of $e^{-0.01} = 0.99$ times. For Bulk, if we switch

the category of bulk from 0 to 1 while holding all other coefficients constant, then the intercept of the log price will decrease 0.103, which means the original price per gram will incrase by a factor of $e^{-0.103} = 0.902$ times. For the confidence interval of all fixed effect variables, all of them do not include 0 , which means they truely have impact on price of Morphine.

For random effect of the state, it means in different states, the base price of Morphine will be a little bit different. The across-state variation is 0.015 and the within-state variation is 0.82, which means the across-state random effct is not strong compared to within state variation. As we can see from the graph, there is certain degree random effect in the model although their contribution is small. For states like Tennessee, Virginia, and Oklahama, the price of Morphin would be higher. However, for states like Arizona, California, and Neveda, the price of Morphine would be lower.



Above plots are the diagonistic plots for the model. In the first plot, all of the points are nearly randomly distributed around the 0 line except there is a small pattern. However, that pattern is acceptable. The QQ plot shows the target variable is deviated fromm the normal distribution. That deviation is expected because the distribution of the $log(price)$ is not that normally distributed as shown in the EDA. For the scale-location plot, there is not an obvious pattern in the graph, which means the variance of the residual is constant across all level of predictions. In the last plot, there are also not any influential outliers exist.

Thus our model is good. At this point, we successfully train a mdoel that can predict the Morphine price.

Limitation

First, when we are exploring the models, we use the forward stepwise selection with the order based on logical reasoning of the relevance between the predictors and the response variable. Since it is forward stepwise selection rather than a method which exhausts all the possible subsets of the predictors, we might miss the optimal combination of the predictors. However, this approach does protect against overfitting since we only consider models which are plausible, so it is a necessary tradeoff. When we built the model, we also cared about the interpretability of model, so we actively avoid the meaningless interactions like the interaction between the `primary reason` and `source`, whose physical meanings were hard to interpret how it was related to the price of the morphine.