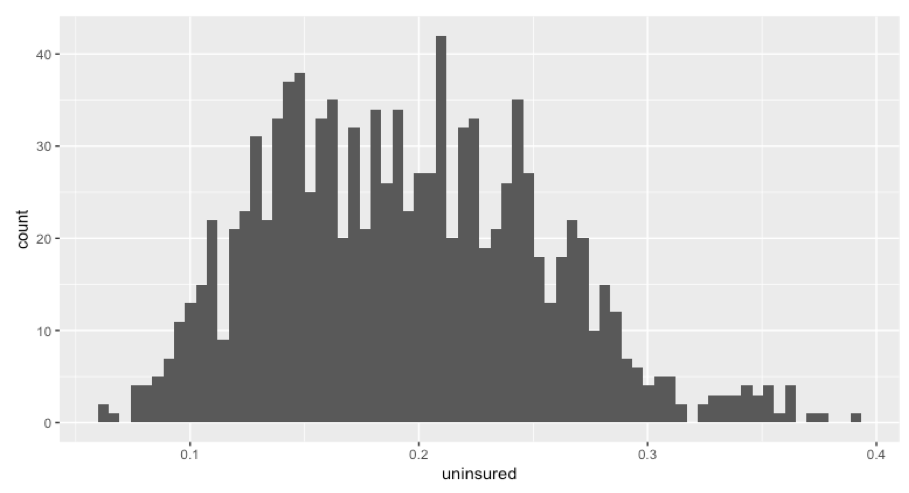
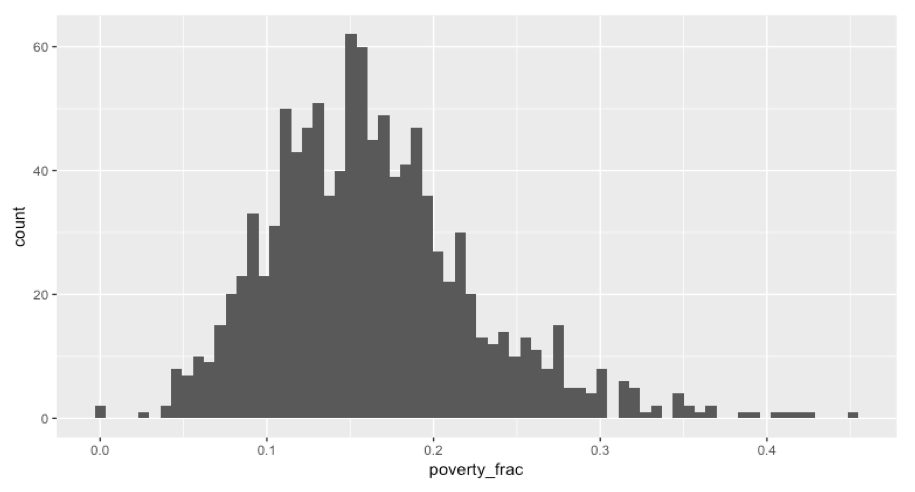
Part 1

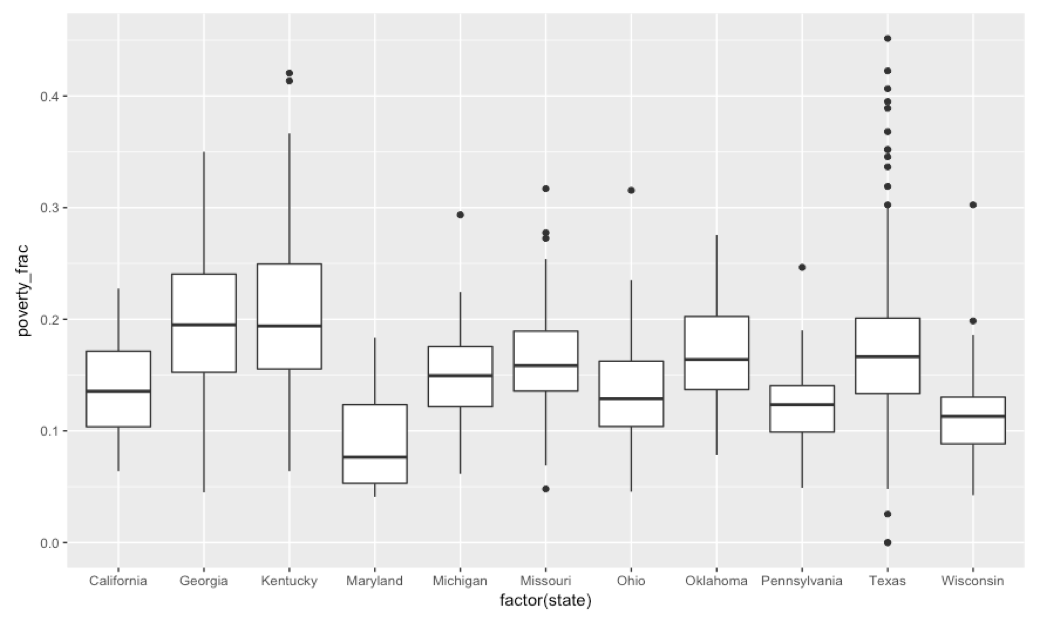
1. 11 states
2. 

It shows that the % of uninsured people are on the low side amongst counties, which might affect their political views. Most counties show that their average uninsured % is around 15% to 23%, roughly speaking.



Similarly, this shows that the % of uninsured people are on the low side amongst counties, which might affect their political views. Most counties show that their average uninsured % is around 10% to 20%, roughly speaking.

1. Here is the boxplot:



As you can see, the boxplots have clear differences in the fraction of people in poverty. For example, Maryland’s box’s lower end is still above Kentucky’s box’s upper land. Also, Texas has a lot of outliers on the upper side, where a lot higher % of people live under poverty, compared to other states. So yes, wealth distribution varies among states.

Part 2

* 1. Yes, since the F-value of the median\_income is 12.839 and the corresponding probability very low at 3.088\*10^(-6).
  2. The default coefficient refers to the high median income. If people have low median income, they are **less** likely to vote for the democratic candidate by 2.74% than if they had high income. Also, if they have medium income, they are less likely to vote for the democratic candidate by 5.36% than if they had high income. Lastly, the fraction of people who vote for the democratic candidate will increase by 2.62% (the difference between the coefficients) if they have low median income compared to the fraction they would have voted for the democratic candidate if they had medium income.
  3. Median\_income is still significant at 0.05 level because it has a very low p-value from the Anova (Type II test), even after controlling for the state.
  4. After controlling for state, if people have low median income, they are **more** likely to vote for the democratic candidate by 2.30% than if they had high income. And if people have medium income, they are **less** likely to vote for the democratic candidate by 1.69% than if they had high income. Finally, if people have low income, they are more likely to vote for the democratic candidate by **3.9895**% than if they had medium income.

The numbers are different from #1 because we are controlling for state. This is because the income and state correlate and we didn’t adjust for collinearity in the first model.

* 1. County shouldn’t be included because it’s a distinct categorical variable that has nothing to do with how people vote. Also, it is all different so it would have just as many different categories as there are data points. Rep12\_frac should not be included because it is the same concept as the response variable. It would perfectly correlate, so it doesn’t add anything to the model.
  2. Non-zero coef variables are:

"(Intercept)" "total\_votes\_12" "registered\_voters"

"reg\_vote\_frac" "turnout" "hs\_diploma\_degree"

"median\_incomeLow" "median\_incomeMedium" "poverty\_frac"

"caucasian\_frac" "gini\_coefficient" "median\_age"

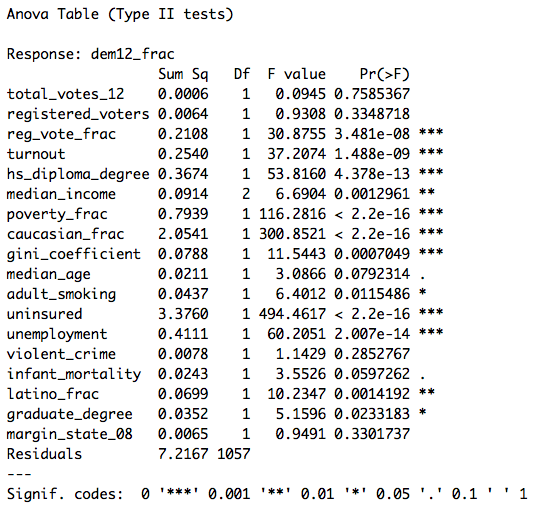
"adult\_smoking" "uninsured" "unemployment"

"violent\_crime" "infant\_mortality" "latino\_frac"

"graduate\_degree" "margin\_state\_08"

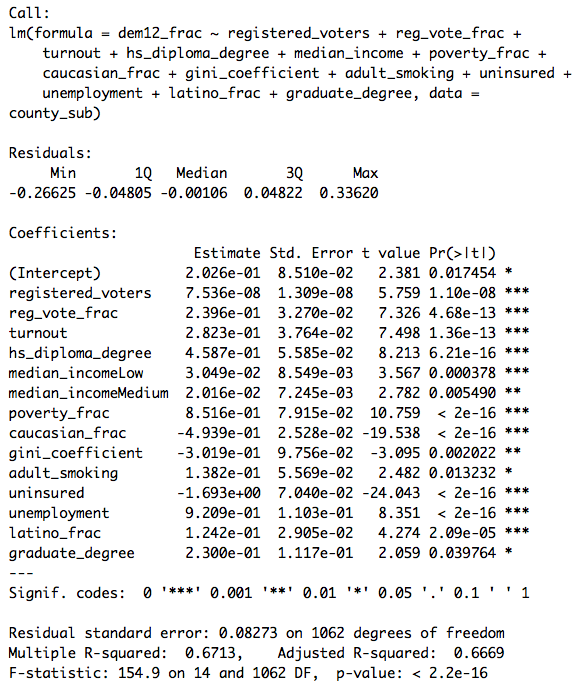
Note that median\_income is a categorical variable.

* 1. As you can see,

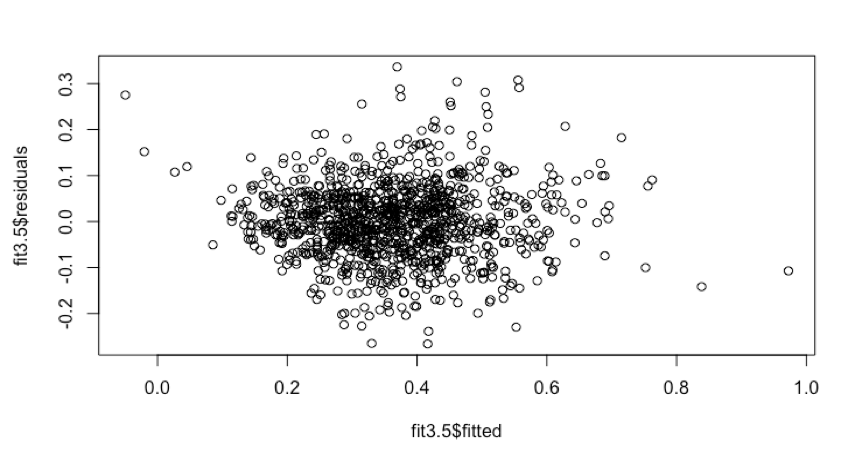


Not all variables are significant at 0.05 level since some of the probabilities are higher than 0.05.

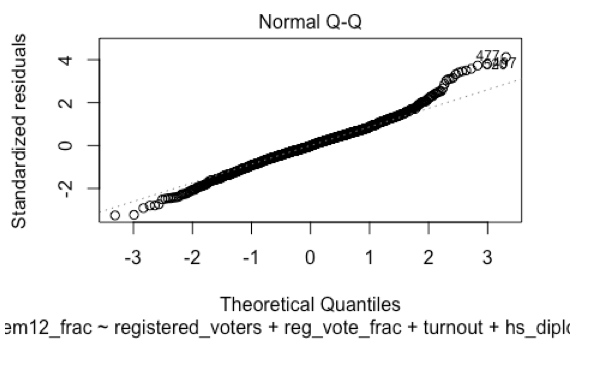
* 1. Summary output below:



* 1. Residual plot below:



Qqnorm plot below:



* 1. For additional 1 registered voter, or 100% of reg\_vote\_fraction, or 100% turn out, or 100% hs diploma, or 100% poverty fraction, or 100% Caucasian fraction, or 100% adult smoking, or 100% uninsured, or 100% unemployed, or 100% latino fraction, or 100% graduate degree,

the fraction of the people in the county that vote for the democratic candidate is affected by

7.536e-08, or 2.396e-01, or 2.823e-01, or 4.587e-01, or 8.516e-01, or -4.939e-01, or 1.382e-01, or 1.693e+00, or 9.209e-01, or 1.242e-01, or 2.300e-01, respectively.

My concern is that even though these are all significant, some of them have marginal impacts such as the impact of one registered voter is 7.536e-08, which is statistically significant but not realistically significant.

This model has 13 variables, which is a lot. When trying to predict, it might be time-consuming to collect all the data for 13 variables. Also, some data such as poverty fraction and uninsured are skewed, because they are on the low side. Also, we didn't include State for simplicity, but adding this in might improve the model and we might be able to get rid of one or more variables that correlate with state, thus making the model more concise.

It seems like the normality assumption is met, based on the qqnorm plot. Also, the residuals look randomly distributed, without any signs of concerns such as heteroscedasticity.

Part 3

* 1. No, because the F-value is small and the corresponding probability is 0.1674, much higher than 0.05.
  2. We can use the model constructed above. Using the model, we get the interval (0.4277473, 0.826941) for the 95% confidence interval.

fit lwr upr

0.6273442 0.4277473 0.826941