What demographic factors are significant in explaining the results of the 2012 and 2016 US Presidential Elections at a county level?

Abstract:

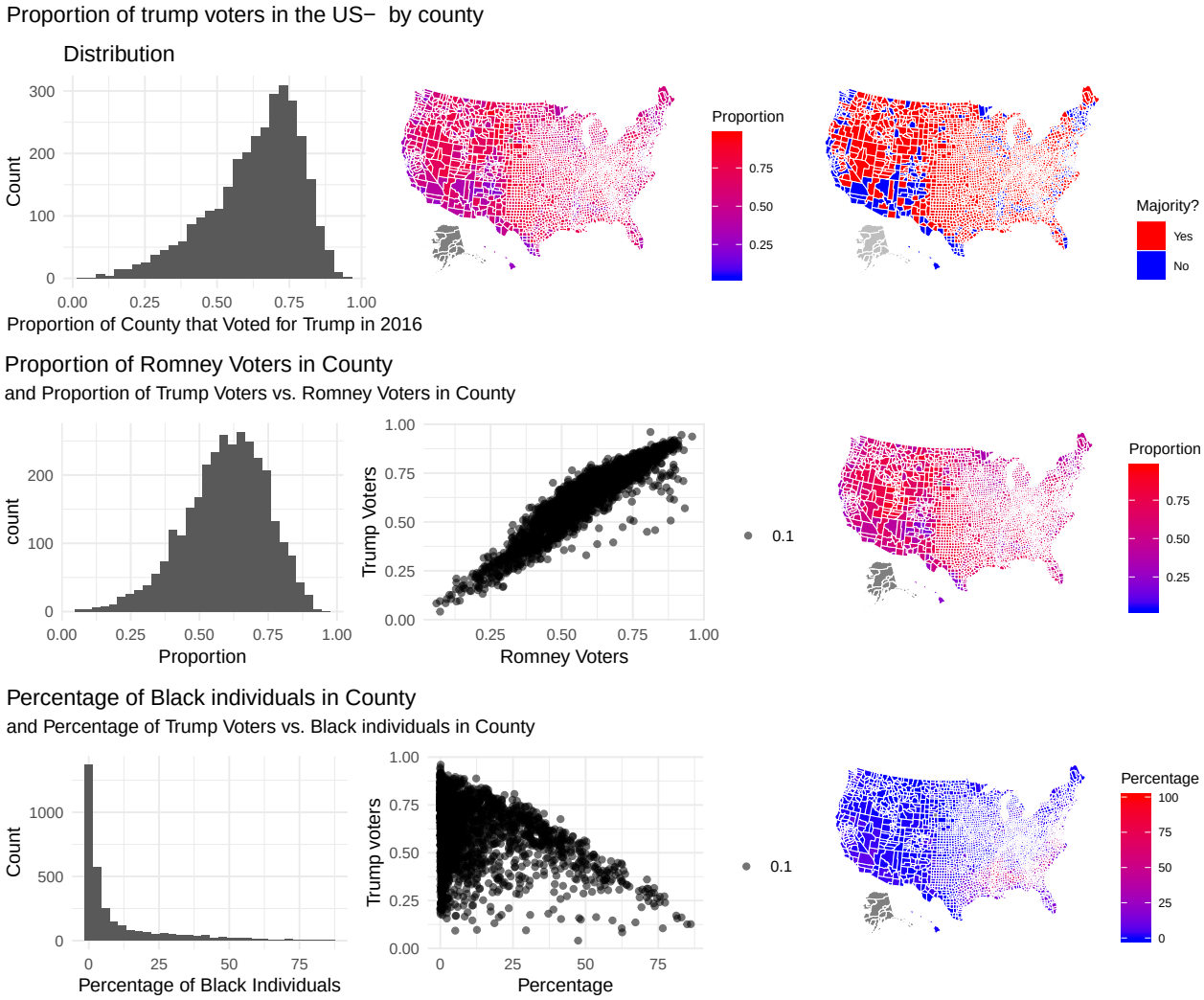
Explaining the results of the 2016 U.S. Presidential Election is beneficial to understanding the way demographic and geographic factors impacted the result. Our model combines multinomial and logistic regression models with a spatial error and lag term. We found multiple demographic factors impacted the proportion of Trump voters in a county, such as the proportion of Romney voters in 2012 and the proportion of individuals with less than a Bachelor's degree, as well as significant spatial autocorrelation. This model is helpful in better understanding the significant factors of U.S. elections and can be adjusted to possibly predict upcoming elections.

Background and Significance

With the 2020 US presidential election is rapidly approaching, it is certainly of interest to determine which demographic factors are most significant in county voting patterns. An article by Citylab states that "it is a fair statement to say... that [demography] favors the democrats". While it would be difficult to predict which factors would be significant in future elections, it would be helpful to better understand Trump's victory in the 2016 election by constructing a retrospective model. We aim to use county-level data from the 2016 election, coupled with demographic information to construct this model.

Methods

We first qualitatively decided which variables would best help us understand the demographic factors behind the proportion of citizens that voted for Trump by county. We did so by choosing variables that were relevant in the news and in prior election coverage. We then performed exploratory data analysis on these variables, looking at how they impacted the 'trump_trans' response variable, our variable that calculated the proportion of Trump voters in each county.



Following this, we used backward selection with adjusted R-squared and BIC as selection criteria. After we arrived at our final model, we considered all possible interaction terms between these predictor variables

and used backward selection with BIC as our selection criteria. Moreover, as many of our variables are demographic categories, we wanted to check for multicollinearity. We did so by looking at the Variance Inflation Factors (VIF) test. We found that the percentage of white individuals in a county and the proportion of Romney voters were rather collinear, so we decided to remove our white_pct variable. Lastly, we wanted to account for violations in the independence assumption for our model, and since our data is aggregated at the county level, we decided that it would be appropriate to perform basic spatial autocorrelation analysis. We then decided to check for spatial autocorrelation in the model using Moran's I. After testing three models, one with the lag term, one with the error term and one with both, the model with both a spatial lag and a spatial error term had the highest log likelihood and a Moran's I value closest to 0. Therefore, our final model includes the main effects from the first backward selection process, the significant interaction effects from the second backward selection process and the spatial lag and error term to account for spatial autocorrelation. The full model can be found in the additional works section.

We found that our model had an R^2 value of 0.9523267, which suggests that around 95% of the variation in the proportion of a county that voted for Trump can be explained by the predictor variables in our final model. Since the value of 95% is relatively high, we know that our findings on significant variables explain the results of the 2016 election relatively well and therefore these demographic factors could be important to observe for future elections. However, this calculation does not mean that our model can explain 95% of the variation for future elections.

Our checks for assumptions can be found in additional work.

Results

Our goal for this project was to understand which variables are most important in analyzing election results. We chose to look specifically at which variables were significant in predicting the proportion of Trump voters in the 2016 election.

We found that some important variables in explaining the proportion of Trump voters by county in 2016 are: the proportion of Romney voters from the 2012 election, percentage of Black individuals, percentage of unemployed individuals, percentage of those with less than a Bachelor's degree, percentage of White individuals with less than a high school diploma, the percentage of the county that is rural, the percentage of the population that is 29 years old or younger as well as the interaction terms between the proportion of Romney voters and the percentage of those aged 29 or younger and between the percentage of the county that is rural and the percentage of those aged 29 or younger.

Based on this model output, the most impactful variable in our model was the proportion of Romney voters. This makes sense as previous party affiliations are usually stagnant and can be relatively accurately predicted on a county level. Therefore, we can conclude that individuals are not likely to change their political affiliations between consecutive elections. Another interesting result was the significance of education level, specifically looking at those with less than a Bachelor's degree. Our model suggests that counties with higher proportions of a lower education level were more likely to vote for Trump. Finally, it is interesting to note how both significant interaction terms include the proportion of individuals of 29 years old or younger in a county. This suggests that the younger individuals were more susceptible to voting trends around them. We found that assuming the number of Romney voters in a county stays constant, an increase in the proportion of 29 or younger is correlated with a decrease in proportion for Trump. We are assuming that voting trends are not going to change massively, and therefore the results from our model can help us understand what might be important in the upcoming election.

Another important result from our analysis is that the 2016 election was affected by spatial autocorrelation correlation. Therefore, in future elections it is important to consider how neighboring states and counties might affect voting results. This is especially relevant to presidential primaries, a topic we did not look at in our analysis, since different states vote on different days. Nonetheless, as we have found, this is also something to consider for the general election.

Discussion and Conclusions

According to our analysis, the significant predictor variables we found in our model are important for analyzing the 2016 presidential elections. Therefore, we think that previous voting patterns, race proportions, ruralness, education levels and age breakdown in a county are crucial demographic factors in explaining the proportion of Trump voters by county. Some counties had very little data available: Bedford, Virginia, Oglala Lakota, South Dakota and Kansas City, Missouri. We looked into these counties, and there does not appear to be any commonalities between them, implying there would be no change in bias in our model.

However, these are predictor variables from our limited dataset and do not reflect all factors influencing the election such as industrial makeup of a county's employment and economy. We have not accounted for determinants of voter turnout, instead only focusing on the proportion of voters that voted for Trump. Looking at issues important to voters for that election cycle may also help us understand voter turnout. This would also help provide a more thorough model that would aim to analyze which demographic in the county will vote. Another drawback to our analysis is that it is constrained by our limited technical knowledge and statistical background to work with, interpret, and draw inferences from spatial data and spatial autocorrelation. While we used the spatial lag model, perhaps more advanced techniques would be appropriate.

If we were to redo the project, we could look for an additional dataset that provided some of these demographic factors that were unavailable to us and combine it with the one we are currently using. Furthermore, we could do extensive research into spatial autocorrelation and how to address it properly, as different methods could help us better explain the geographical relationship between counties and the proportion of Trump voters. Similarly, we could also try to determine how the likelihood that an individual would vote at all would impact the proportion of Trump voters in a county, as this could also depend on demographic and geographic factors.

It would also be interesting to change this model to try and predict the proportion of Trump voters by county for the upcoming 2020 presidential elections. This could be done by using the current data as a training set for K-fold cross validation to create the best model and then use the most recent demographic data as the testing set for the upcoming elections. Thus, it would be beneficial to us to gain access to data past 2016 if we were to improve this model to identify the proportion of Trump voters in each county for the 2020 election. Our 2016 model suggests that previous party affiliation, college education levels and age demographics are influential in election results.

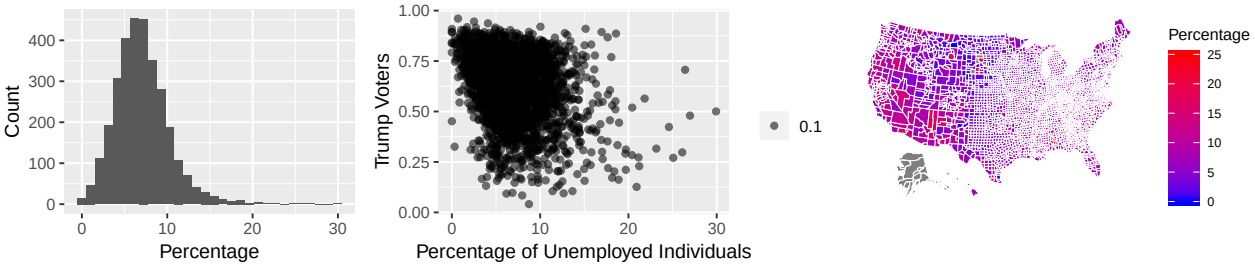
References

Misra, T. (2016, February 26). Demographics Explain the Likely Outcomes in the 2016 U.S. Presidential Election. Retrieved June 25, 2020, from https://www.bloomberg.com/news/articles/2016-02-26/demographics-explain-the-likely-outcomes-in-the-2016-u-s-presidential-election'

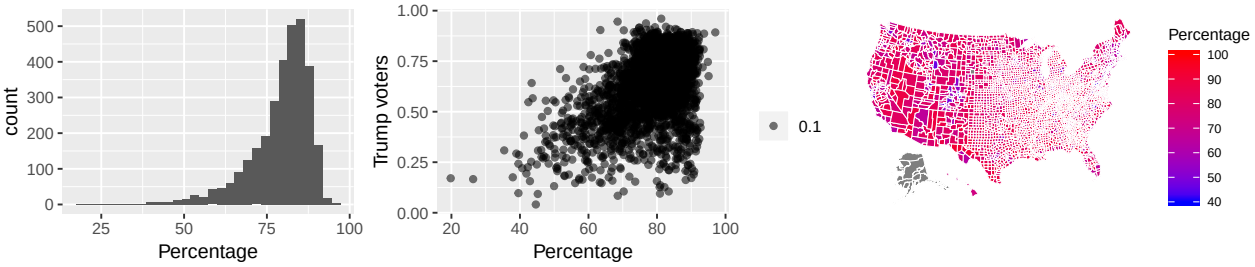
Additional work

Additional Exploratory Data Analysis

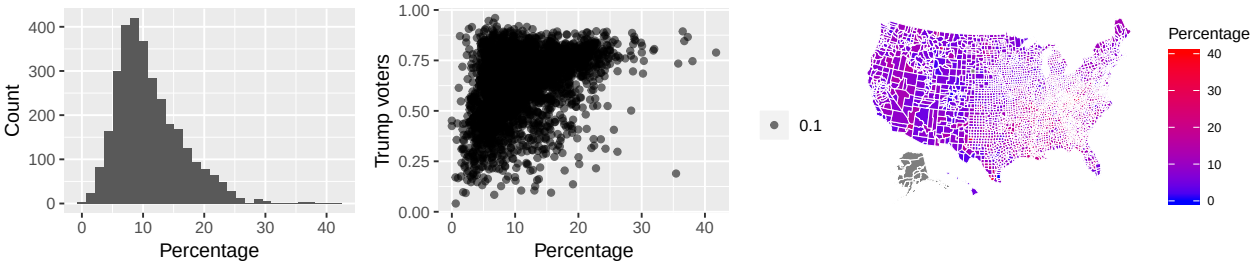
Percentage of Unemployed Individuals in County
and Proportion of Trump Voters vs. Unemployed In County



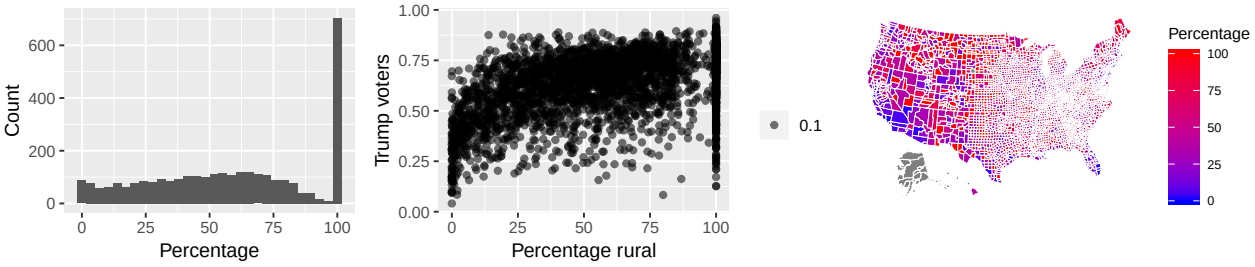
Percentage of Population with Less Than Bachelor's Degree in County
and Proportion of Trump Voters vs. Less than Bachelor's Degree In County



Percentage of White individuals with Less Than High School Diploma in County
and Proportion of Trump Voters vs. White individuals with Less Than High School Diploma In County



Percentage of Rural Population in County
and Proportion of Trump Voters vs. Rural Population In County



Here is our selected model:

```
## # A tibble: 10 x 5
##    term                          estimate std.error statistic   p.value
##    <chr>                            <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                     -2.98     0.0960     -31.0  6.38e-184
##  2 romney_prop                      5.27     0.145       36.4  1.87e-241
```

```
##  3 black_pct                      -0.00464  0.000263     -17.6  2.29e- 66
##  4 clf_unemploy_pct               -0.0145   0.00128      -11.3  5.85e- 29
##  5 lesscollege_pct                 0.0135   0.000516      26.1  4.02e-136
##  6 lesshs_whites_pct               0.00909  0.000851      10.7  3.68e- 26
##  7 rural_pct                      -0.00482  0.000697      -6.92 5.50e- 12
##  8 age29andunder_pct               0.00384  0.00226        1.70 8.88e-  2
##  9 romney_prop:age29andunder_pct -0.0371   0.00368      -10.1  1.77e- 23
## 10 rural_pct:age29andunder_pct    0.000173 0.0000185      9.35 1.62e- 20
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl>  <dbl>
## 1     0.942         0.942 0.176     5583.       0    10   993. -1964.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

We looked at the Variance Inflation Factors (VIF)

```
## # A tibble: 9 x 2
##   names                           x
##   <chr>                       <dbl>
## 1 romney_prop                 46.3
## 2 black_pct                    1.46
## 3 clf_unemploy_pct             1.72
## 4 lesscollege_pct              2.23
## 5 lesshs_whites_pct            2.06
## 6 rural_pct                   48.1
## 7 age29andunder_pct           15.1
## 8 romney_prop:age29andunder_pct 53.0
## 9 rural_pct:age29andunder_pct   40.7
```

based on VIF- we removed white_pct

Combined spatial error and spatial lag model:

```
##
## Call:
## sacsarlm(formula = trump_trans ~ romney_prop + black_pct + clf_unemploy_pct +
##     lesscollege_pct + lesshs_whites_pct + rural_pct + age29andunder_pct +
##     romney_prop * age29andunder_pct + rural_pct * age29andunder_pct,
##     data = merged, listw = ww, zero.policy = T)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.2161193 -0.0889066  0.0061097  0.1011747  1.5023333
##
## Type: sac
## Coefficients: (asymptotic standard errors)
##                                  Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)                   -2.9776e+00  9.6149e-02 -30.9689 < 2.2e-16
## romney_prop                    5.2669e+00  1.4460e-01  36.4232 < 2.2e-16
## black_pct                     -4.6407e-03  2.6277e-04 -17.6607 < 2.2e-16
## clf_unemploy_pct              -1.4486e-02  1.2820e-03 -11.2998 < 2.2e-16
## lesscollege_pct                1.3488e-02  5.1548e-04  26.1661 < 2.2e-16
## lesshs_whites_pct              9.0922e-03  8.4992e-04  10.6977 < 2.2e-16
## rural_pct                     -4.8160e-03  6.9578e-04  -6.9217 4.462e-12
## age29andunder_pct              3.8597e-03  2.2537e-03   1.7126   0.08679
## romney_prop:age29andunder_pct -3.7080e-02  3.6754e-03 -10.0887 < 2.2e-16
```

```
## rural_pct:age29andunder_pct    1.7313e-04  1.8504e-05   9.3566 < 2.2e-16
##
## Rho: 0.00035445
## Asymptotic standard error: 0.0016562
##     z-value: 0.21402, p-value: 0.83053
## Lambda: -0.00076777
## Asymptotic standard error: 0.0055072
##     z-value: -0.13941, p-value: 0.88912
##
## LR test value: 0.054611, p-value: 0.97306
##
## Log likelihood: 992.8725 for sac model
## ML residual variance (sigma squared): 0.030925, (sigma: 0.17586)
## Number of observations: 3111
## Number of parameters estimated: 13
## AIC: -1959.7, (AIC for lm: -1963.7)
```

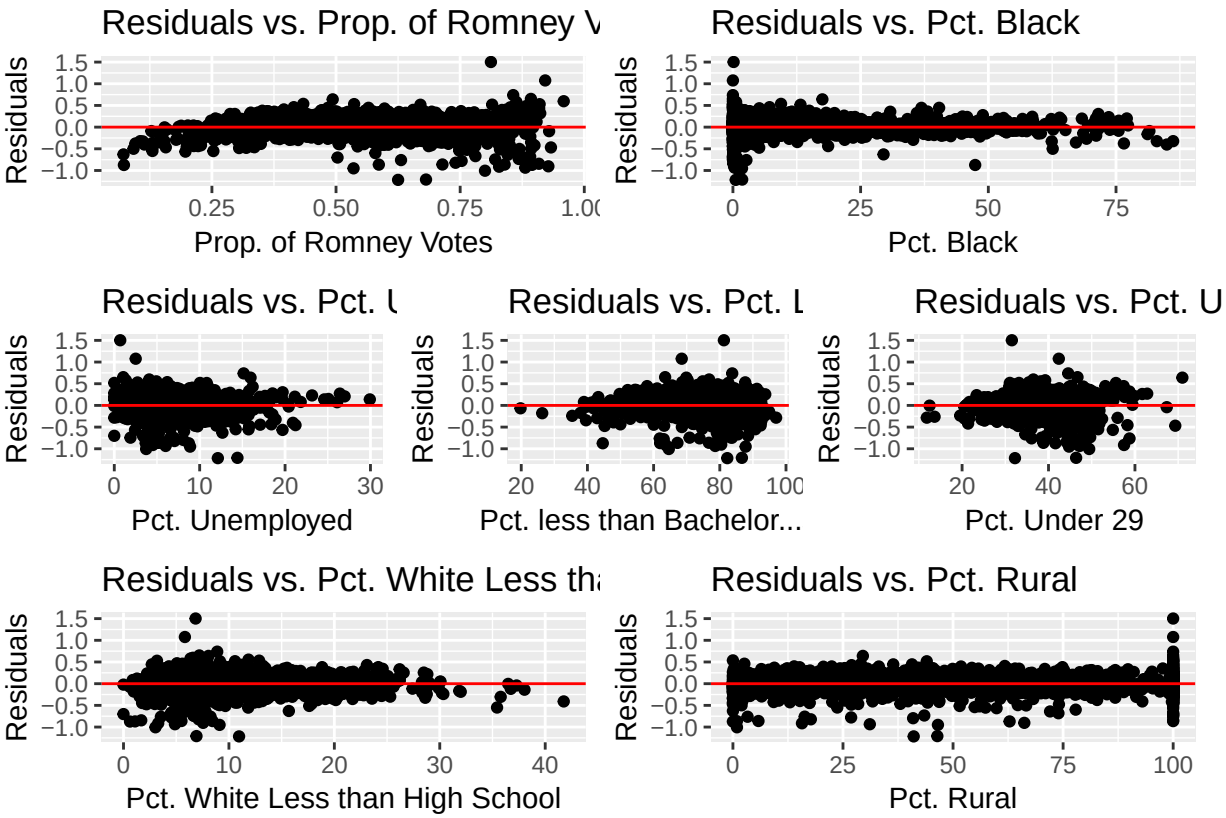
Moran's I test:

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  merged$residuals_both
## weights: ww
## number of simulations + 1: 1001
##
## statistic = -5.44e-06, observed rank = 529, p-value = 0.4715
## alternative hypothesis: greater
```

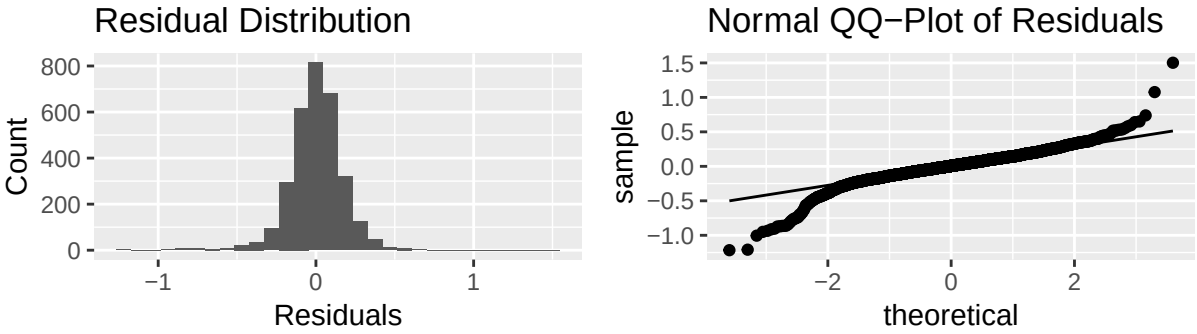
Assumptions:

When checking the assumptions we plotted the residuals vs predicted/ individual predictors. Based on our residual graphs (which can be found in additional work), there seems to be some fan shaped distributions of residuals, so we will have to proceed with caution with regards to linearity. Likewise, for constant variance, the fanning pattern shows a slight violation. Therefore, we will proceed with caution. Since we applied the logit function to the proportion of Trump voters, we were able to calculate a response variable that can take all possible values between negative and positive infinity. We also graphed a histogram of our residuals (additional work), and found that the distribution of our residuals is very unimodal and symmetric around zero which satisfies our normality condition. For most of the normal-QQ plot, it follows a straight diagonal line, which also shows that our model satisfies the normality condition. However, it is important to note that the normal-QQ plot deviates from the line around the ends, so we should proceed with slight caution. As stated in our introduction, there seems to be strong signs of spatial autocorrelation in the 'trump_trans' response variable, which is a clear violation of independence. We tried to account for spatial autocorrelation in our analysis, since it is inherently geographic data. As mentioned above, we tried a spatial lag and error term to account for this in our final model. Due to the spatial lag model having a lower AIC and Moran's I value closest to zero, we chose to proceed with this model.

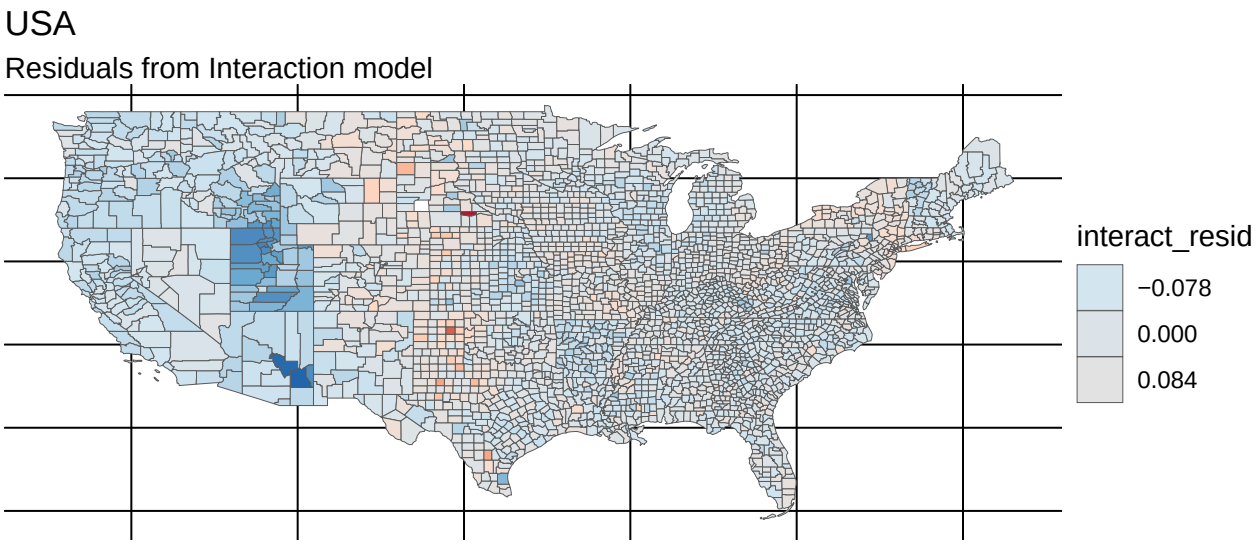
Linearity graphs:



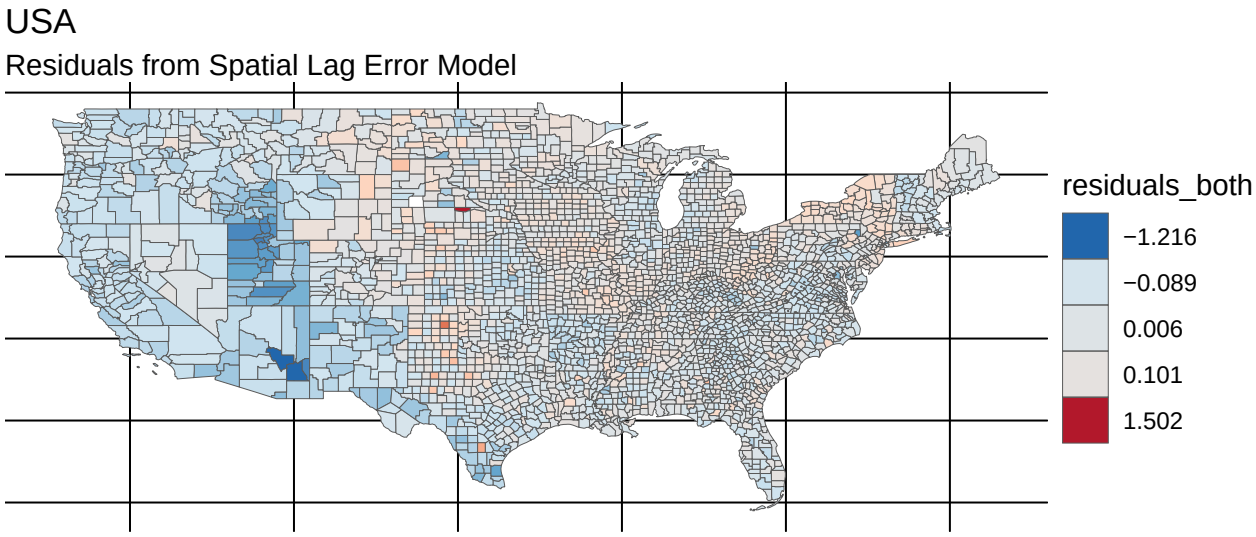
Normality graphs:



Independence plots:



prior to model



post model