Trinity College Dublin

MSc Computer Science-Data Science Strand

CS7DS3 Applied Statistical Modelling

Main Assignment

Daim Sharif – 24345724

30/04/2025

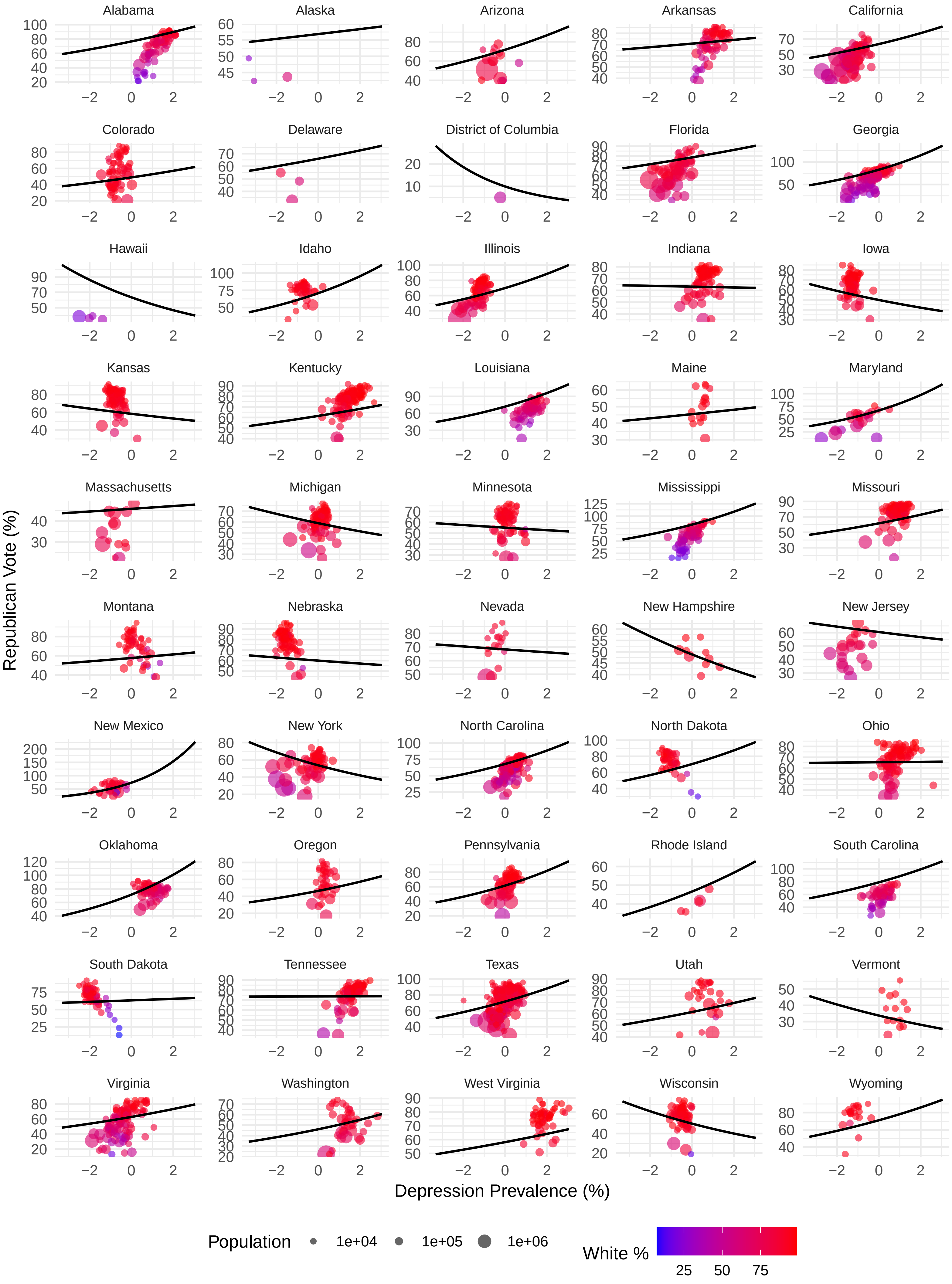
Objective Statement

1. I investigated whether depression prevalence was associated with increased support for Donald Trump in the U.S. presidential election.
2. Initial methods like simple linear regression, random slope and intercept models with depression only, showed a statistically significant positive association between depression prevalence and Republican votes.
3. However, once I added other factors like key demographics, including race, gender, and population size, the association between depression and Republican support weakened and, in some models it lost significance.
4. Factors related to demographics including racial composition, gender, population size came out stronger and more consistent predictors of republican vote share.
5. Mixed-effects models with random slopes for depression by state showed variation in the association between per_gop and depression across states. Where in some states, depression predicted more Republican voting, while in others, the relationship was weak or even reversed.
6. Alternative modelling strategies were experimented where I added non-linearity and interaction terms in the random slope fit. Though there were some improvements in the fit, it did not fundamentally change the finding that depression was not a dominant predictor of Republican support at the state level.

The central figure on the next page shows the relationship between depression prevalence and Republican vote share at the county level, broken down by state. It was modelled using the random slope model, which models the log of Republican vote share as a function of depression prevalence, racial composition, gender ratio, and total votes, while allowing the depression effect to vary by state. The population of each county is encoded using size and the racial composition is encoded using colour. Looking at the figure we can see that depression prevalence is centered, meaning each state's trend line shows the effect of being above or below the national average. The figure shows that county-level depression is linked to Republican voting in some parts of the country and that these relationships are intertwined with demographic composition and population size. For instance, in states like Oklahoma, Alabama, Wyoming, Texas, and South Carolina, counties with higher depression rates tend to have higher GOP vote share. These panels show many red points, clustered top-right showing more depressed counties has higher GOP share. Whereas, in states like New York, where we can see larger points, bigger population and bluer colours, lower white race composition, showing counties where higher depression does not translate to more Republican voting.

Depression Rates vs. Republican Voting by State

Mixed−effects random slopes per state; point size = population, color = % white



Project Workflow and Analytical Approach

1. Data Preparation and Exploratory Data Analysis

For this assignment we had to conduct analysis using a dataset containing U.S. county-level data on voting outcomes from the presidential election, alongside demographics and health indicator. The primary response variable of interest for the proportion of votes cast for the Republican candidate (Donald Trump), calculated as a percentage (per_gop) based on the number of votes received out of the total votes cast in each county.

Variable Name	Description
STNAME	Name of the U.S. state
CTYNAME	Name of the county
TOT_POP	Total population of the county
TOT_MALE	Total number of males in the county
TOT_FEMALE	Total number of females in the county
total_votes	Total number of votes cast in the U.S. presidential election
per_gop	Percentage of votes for the Republican party (Donald Trump)
Crude.Prevalence.Estimate	Estimated percentage of adults reporting frequent mental distress (depression)
race	Proportion of residents identifying as white and not any other race

Table 1: Variables in dataset

Table 1 shows the variables present in the dataset. To have a better know about the variables I plotted a covariance plot which shows the correlation between the pair of variables, their trend via a scatter plot, and the distribution of the variables.

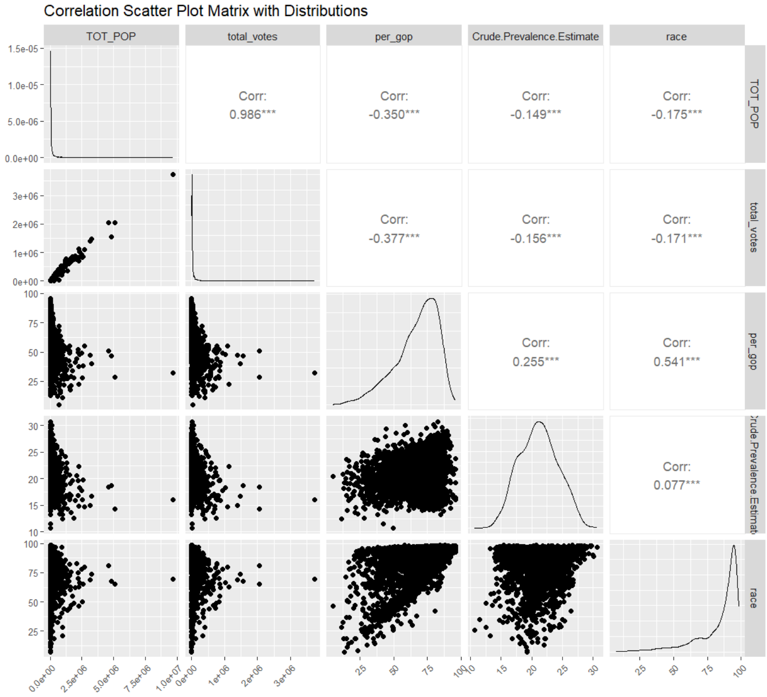


Figure 1: Correlation scatter plot with distributions

Figure 1 shows the correlation scatter plot with distributions for the continuous variables. In terms of the distribution, we can see that TOT_POP and total_votes are highly skewed, followed by race, and per_gop. This can affect the results in our modelling, indicating need of transformation. Furthermore, amongst the correlations, we can see that crude.prevalance. estimate and per_gop have a positive correlation, also seen by the scatter plot, although not clear. Nevertherless, we are interested in state level results for the core part of our findings.

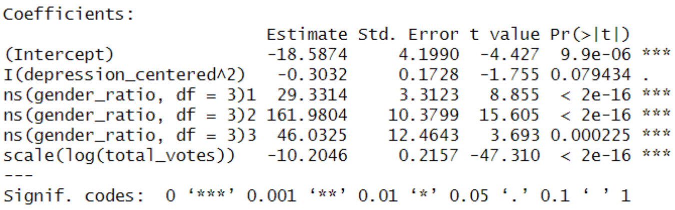
Before I could start modelling, I had to transform these variables suited for modelling. Refer to table 2 for a summary of these transformations.

Transformed Variable	Original Variable	Transformation Applied	Purpose / Description
log_per_gop	per_gop	log(per_gop)	Stabilize variance and interpret % change in vote share multiplicatively.
log_total_votes	total_votes	Scale(log(total_votes))	Normalize skewed voting count data for modeling. With stabilize variance and centring
depression_centered	Crude Prevalence Estimate	scale(Crude Prevalence Estimate)	Mean-center depression for interpretability in mixed models
race_sc	race	scale(race)	Standardize proportion of white population to interpret effect size in units of SD.
gender_ratio	TOT_MALE / TOT_FEMALE	Created + optionally scaled	Constructed variable showing male-to-female ratio in each county.

Table 2: Variable transformation

2. Modelling and Evaluation of Results

Initially I fit a simple linear regression model on per_gop, weighing the parameters with sqrt(TOT_POP), including all the transformed variables and giving an adjusted R^2 value of 60.7%. However, it was too simple for the objective of this assignment. I then added an interaction term depression_centered^2 for non-linearity, and splines for gender-ratio, keeping the rest the same, remember that for the rest of the models I weighed the parameter with sqrt(TOT_POP). This model performed worse with an adjusted R^2 value of 49% but most importantly, it caused the depression variable as seen by figure 2 to become not significant defeating the purpose of this assignment.



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -18.5874     4.1990  -4.427  9.9e-06 ***
I(depression_centered^2)   -0.3032     0.1728  -1.755 0.079434 .  
ns(gender_ratio, df = 3)1  29.3314     3.3123   8.855  < 2e-16 ***
ns(gender_ratio, df = 3)2 161.9804    10.3799  15.605  < 2e-16 ***
ns(gender_ratio, df = 3)3  46.0325    12.4643   3.693 0.000225 ***
scale(log(total_votes))   -10.2046     0.2157 -47.310  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Coefficients after adding interaction terms and splines

As a further robustness check, before we move on to mixed-effect models. I fit a logit-transformed linear modelled including state as a fixed effect. In this approach, the model estimates the per_gop in each state while holding depression and demographics constant. Depression remained a significant predictor, suggesting that its association with Republican voting stays even after accounting for inter-state differences. However, this model assumes the depression effect is constant across states and thus does not capture state level variation which is better addressed by our random slope model. For this model, I got an adjusted R^2 of 77%, although it is higher, I am adding 50+ additional parameters because of the state factor, inflating the fit.

I then moved to mixed-effect models, in other words, random intercept and random slope models. To appropriately account for the hierarchical structure of the data, with counties nested within states, I used mixed effects/ hierarchical models. These models allow to model both the fixed effects within the predictors like depression prevalence, racial composition, gender ratio, etc. and random effects that capture state-level variability. To begin with I first fit a random intercept model on log(per_gop) with depression as fixed effect and STNAME as random effect only which showed a significant association, however, this was an unadjusted association between depression prevalence and per_gop, later when I added the other fixed factors, the effect of depression weakened as I adjusted for cofounders, or variables that are correlated with depression and per_gop, particularly racial composition and total votes (remember that I am using transformed variables throughout the modelling). Certainly, the later model was a better fit through anova evaluation, with it having a lower AIC and BIC as well. I then fitted a random intercept and slope model, allowing the effect of depression vary by state. This approach suited the objective of this assignment. Racial composition and total_votes emerged strong consistent predictors of per_gop, depression prevalence showed a modest but statistically significant positive association with Republican voting as seen in figure 6, but this effect varied across states, as captured by the random slopes, supported by figure 3. The X-values are on log scale since I modelled log(per_gop).

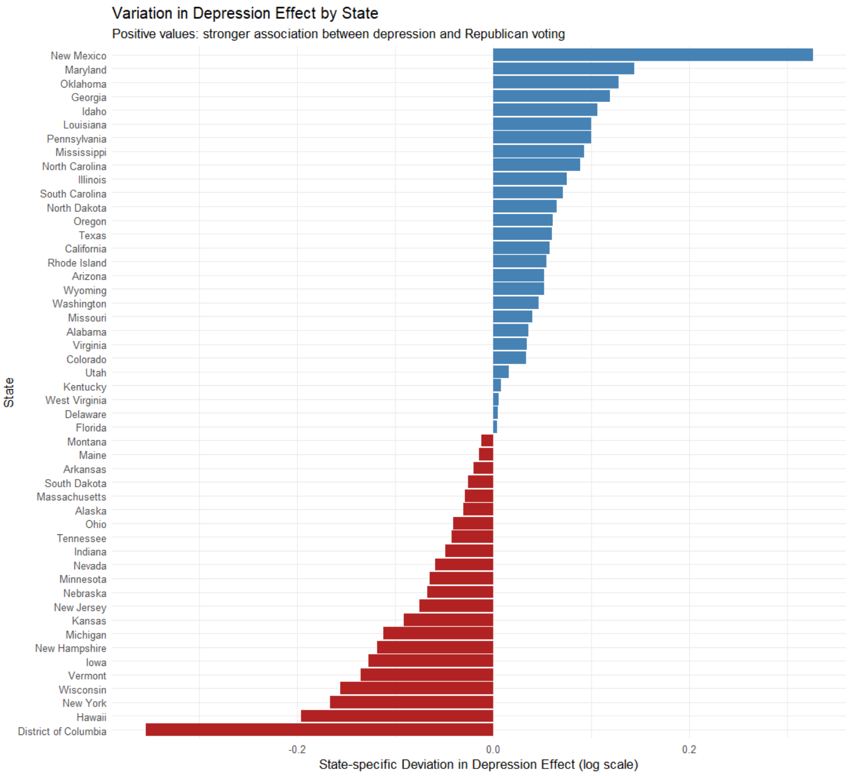


Figure 3: Variation in depression effect by state captured by random slope model

Figure 3 shows that the relationship between depression prevalence and Republican support is not consistent across the US. The geographic heterogeneity is evidence that state context matters and justifies the use of a random slope model. The bars to the right (blue) have a stronger positive association between depression prevalence and Republican voting than the national average, as opposed to the red, which is the opposite, showing a weaker, or even a negative depression-voting relationship.

To improve the performance of the model I started from modelling per_gop and untransformed fixed effects and random effects to modelling log(per_gop) and transformed fixed and random

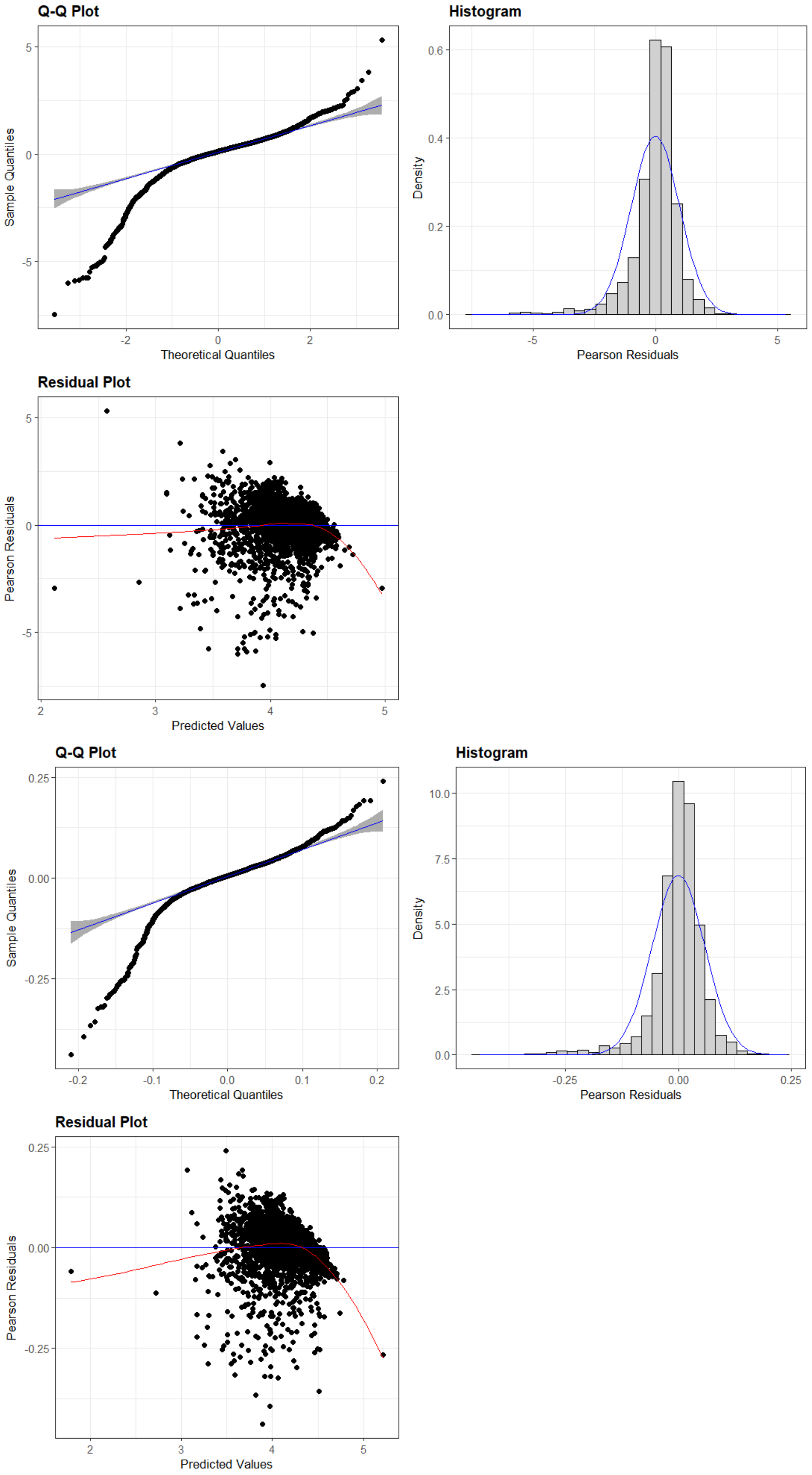
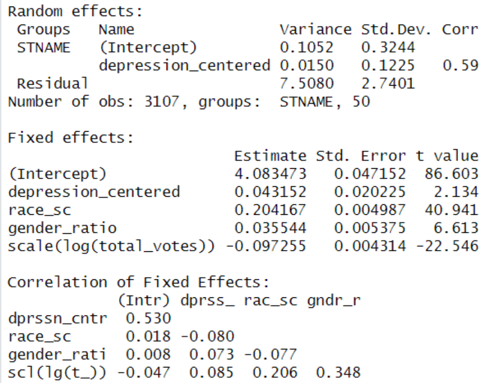


Figure 4: Before (top) and after (bottom) random slope model improvement

effects, along with weighing the parameters, as seen in table 2. Although anova proved that the later model was better. I wanted to test the model assumptions.

I examined the residual diagnostics before random slope model improvement and after. The histogram of residuals after improvement is well centred at 0 and follows the blue density line better, as opposed to before the improvement. Moving on the qq-plots on the top left shows curved deviation in the tails before improvement indicating that residuals are not normally distributed which gets better after the improvement. Finally, the homoscedasticity plot on the bottom left show stronger fanning whereas it is slightly more stable after the performance with the residuals more tightly clustered around 0. Therefore, there is an improvement in the model as it is closer to satisfying the normal and there is less residual spread and better fit. I tried improving the model by adding interaction terms depression*race to examine its relationship hoping to better the model and using a nonlinear spline model for depression to detect potential curvilinear associations. Although, the model slightly improved fit through anova testing and residual diagnostic, the models lacked clear interpretability and did not meaningfully change the conclusions that the depression effect remained weak after adjusting for demographics. Here is the final model log(per_gop) ~ depression_centered + race_sc + gender_ratio +   scale(log(total_votes)) + (1 + depression_centered | STNAME)



```
Random effects:
 Groups   Name                Variance Std.Dev. Corr
 STNAME   (Intercept)         0.1052   0.3244
          depression_centered 0.0150   0.1225   0.59
 Residual                     7.5080   2.7401
Number of obs: 3107, groups:  STNAME, 50

Fixed effects:
                         Estimate Std. Error t value
(Intercept)              4.083473   0.047152  86.603
depression_centered      0.043152   0.020225   2.134
race_sc                  0.204167   0.004987  40.941
gender_ratio             0.035544   0.005375   6.613
scale(log(total_votes)) -0.097255   0.004314 -22.546

Correlation of Fixed Effects:
            (Intr) dprss_ rac_sc gndr_r
dprssn_cntr  0.530
race_sc      0.018 -0.080
gender_rati  0.008  0.073 -0.077
scl(lg(t_)) -0.047  0.085  0.206  0.348
```

Figure 6: Summary of random slope model

The central figure was built using the final random slope model. For each state, I predicted Republican vote shares based on the model's fixed and random effects. Because I modelled using a log transformation on the response variable, predictions were transformed back easily by exponentiating them for easier interpretation. The final model shows depression and per_gop vary across state with random slope variance being 0.015 as seen in figure 6. And because it has a correlation of 0.59 between state-level intercept and depression slope, it suggests that states with higher baseline GOP support show a strong depression-vote link. Furthermore, fixed effect interpretations are as follows in table 3. These trends are supported by the central figure on page 3.

Variable	Effect Direction	Interpretation
Depression (standardized)	Positive	Counties with higher-than-average depression had slightly higher GOP vote share, all else equal.
Race (standardized)	Strong Positive	Counties with a higher proportion of white residents were substantially more likely to vote Republican.
Gender Ratio	Positive	A higher male-to-female ratio correlated with more Republican support.
Total Votes (log-scaled)	Strong Negative	Larger counties (more voters) tended to vote less Republican.

Table 3: fixed effects interpretation

Furthermore, throughout the assignment multiple R packages were used like lme4 for random intercept and slope, ggplot2 for visualization, splines for modelling non-linear effects, performance and ggResidpanel for diagnostics.