CS7DS3 Applied Statistical Modelling

Main Assignment

Harinath Babu (24340502)

Objective statement

The analysis investigates the relationship between depression rates and GOP support, accounting for demographic factors and state level variations of the effect using hierarchical modelling. It shows that higher depression rates are associated with a county having Republican majority in the last election, even after accounting for the other features such as racial proportions. Counties with lower population and higher proportions of white population tend to support the GOP. But state-level effects influence both the baseline support for the GOP, and the effect of depression on the support for GOP with most states amplifying the positive effect of depression while some reverse it. The findings suggest that although a positive relationship between depression levels and support for the Republican party exists on average, regional differences play a vital role in defining the exact relationship.

Exploratory Data Analysis

An exploratory data analysis was conducted to understand the data distributions and the relationships between the features. Then the data was preprocessed. The feature names were cleaned up to be consistent; A new binary feature *majority_gop* was created as 1 if `per_gop > 50` else 0 to denote a Republican party majority; The total population was shifted to a log scale as the total population was skewed; The crude prevalence(depression) estimate, proportion of white population, and (log) total population were standardised using `scale()` as they were all in different ranges to help with the modelling. The state was treated as a factor.
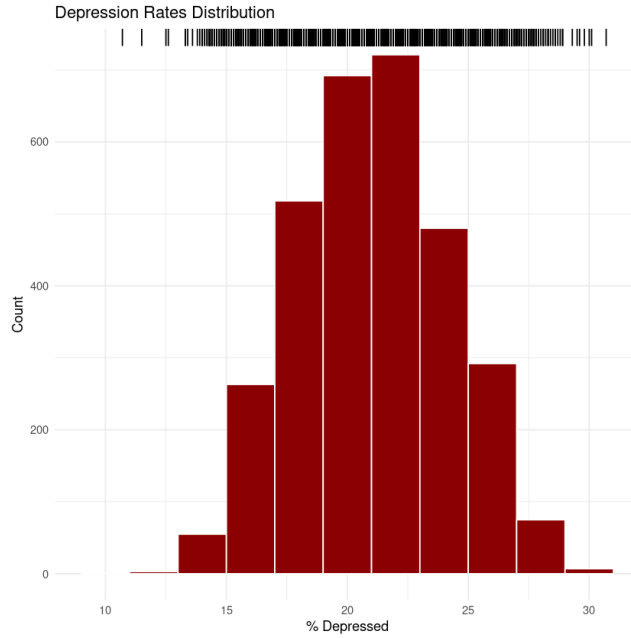
Distribution of Depression rates

Figure 1: Depression Rates Distribution

The histogram of depression rates shows a slightly right skewed distribution, approximately from 10% to 35% peaking around 20%. The overlaid rugplot indicates no apparent outliers, suggesting that the data is suitable for modelling without any preprocessing.

Features Vs GOP majority

(a) Depression Rates   (b) White population %   (c) Log population
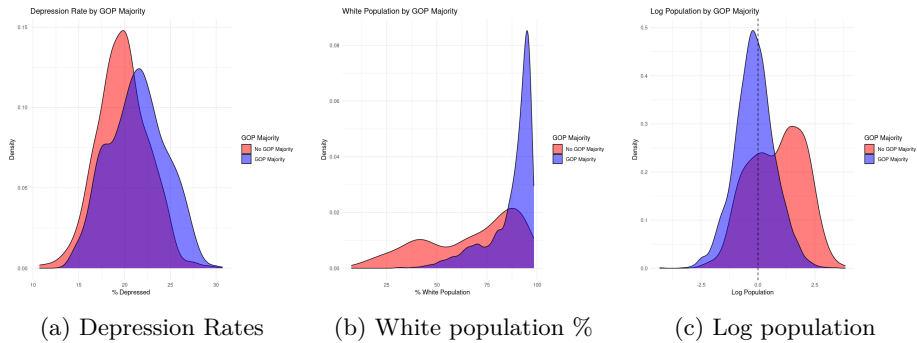
Figure 2: Comparison of key features against GOP majority

The density plot of depression rates by GOP majority status shows a pattern. Counties with a GOP majority tend to have a higher depression rate on average compared to those without a GOP majority. This suggests a potential positive relationship between depression and voting for the Republican party.

The proportion of white population by GOP majority status also shows a clear pattern. Counties with a GOP majority have a higher proportion of white residents. Thus, race proportion is likely a key indicator of voting behaviour which must be controlled for to isolate the effect of depression in the model.

The transformed population plot by GOP majority shows that counties with GOP majority tend to have smaller populations, as indicated by negative scaled log transformed values while Non GOP majority counties are more evenly distributed. This suggests that less populous, rural counties are more likely to be Republican.

Modelling

A Bayesian hierarchical logistic regression model was fitted using the `brms` package. The model predicts majority_gop with fixed effects for depression, race and population, and random effects for state including both a random intercept and a random slope for depression as, `majority_gop ~ scaled_depression + scaled_race + scaled_log_pop + (1 + scaled_depression | state)`. The Bernoulli with the default logit link was used for the binary outcome. Priors were specified as follows: normal(0, 2) for fixed effect coefficients, student_t(3, 0, 2.5) for the intercept, and normal(0, 1) for standard deviations of random effects. These weakly informative priors help regularize the estimates and ensure stable convergence, while still allowing the data to drive inference. Four chains were run, and convergence was confirmed as Rhat($= 1$) and high ESS.

Model selection

Four models were evaluated using Leave one out cross validation; The aforementioned hierarchical model, a hierarchical logistic regression model with the same fixed effects, but only random intercepts for state (assuming common depression effect among states), one that completely disregards depression, and one that models only the random state level intercepts. The results of `loo_compare()` is given below,

Model	ELPD Diff.	SE Diff.
Full model (w/random slopes)	0.0	0.0
Hierarchical model (w/o random slopes)	-14.2	7.0
No depression model	-29.0	9.0
State-only model	-451.3	29.8

The model with state level random slopes for depression shows the best per-

formance. Removal of the random slopes led to a worse performance (ELPD diff = -14.2; 2 SE worse). Excluding depression worsened the performance heavily (ELPD diff = -29.0; 3.2 SE worse) reinforcing the effect for depression, and the state only model was quite worse(ELPD diff = -451.3; 15.1 SE worse) highlighting the importance of the chosen predictors.
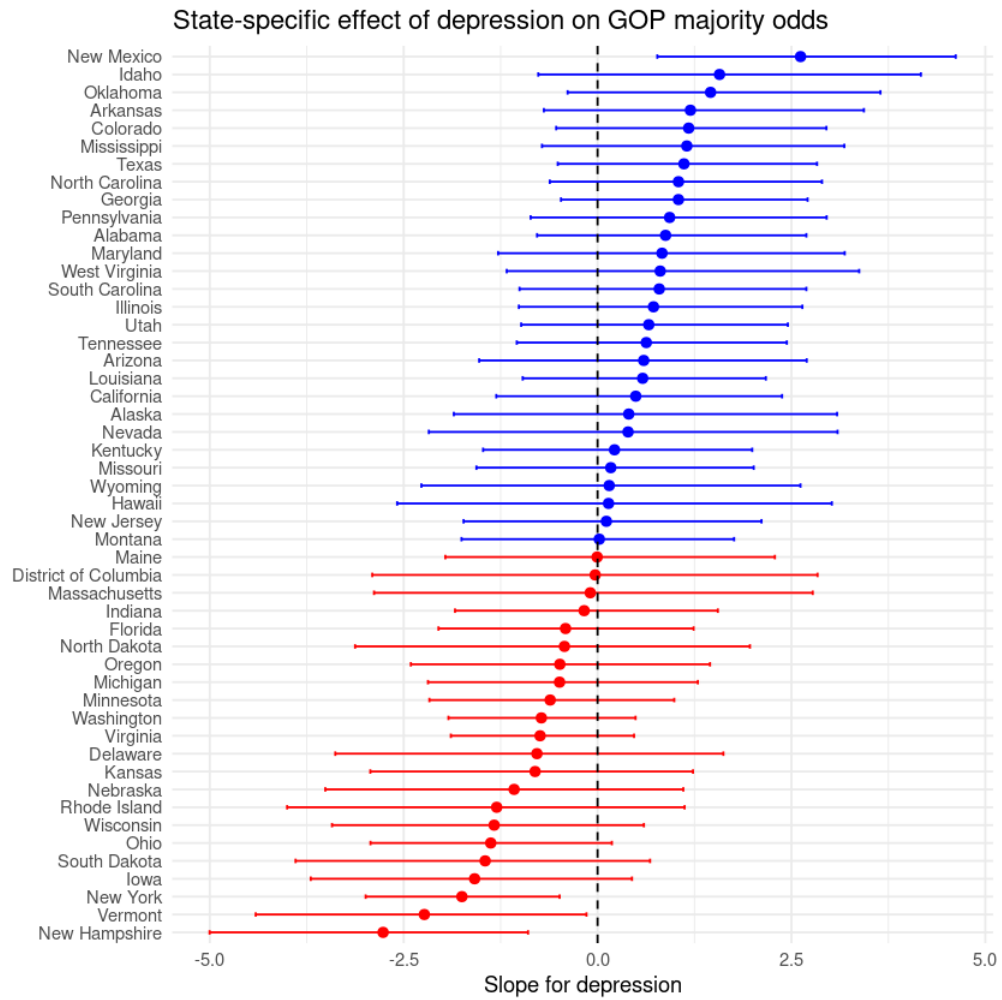
Results

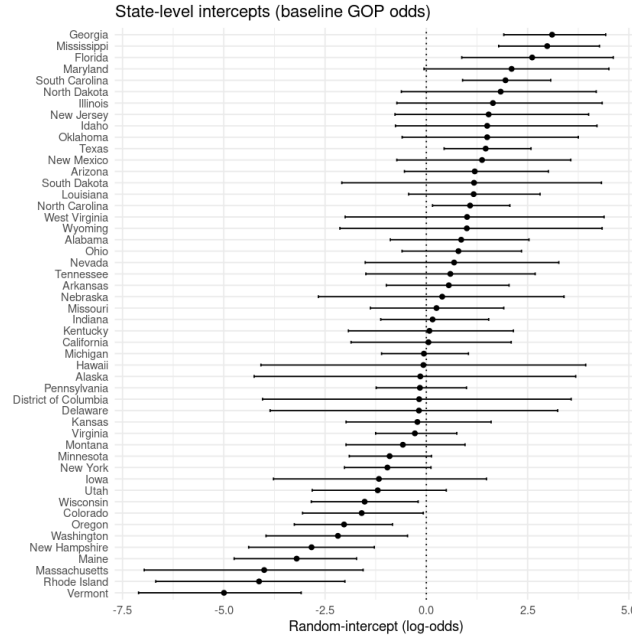Figure 3: Central figure - State specific effect of depression on GOP majority odds

4

Figure 4: State level effects on baseline GOP majority odds

Fixed effects

Predictor	Estimate	95% CrI
Intercept	2.62	[1.93, 3.29]
Depression	1.03	[0.44, 1.61]
% White (Race)	1.87	[1.65, 2.10]
Log Population	-1.05	[-1.23, -0.87]

- The Odds ratios can be obtained by exponentiating the model coefficients, which provide a more intuitive interpretation.
- In an average county (one where all other predictors are at their mean), the baseline odds of a GOP majority is 14 to $1 (e^{2.62} \approx 13.74)$.
- A 1 SD increase in depression rate i.e., by 3.14% (approx.) results in a 2.8 times $(e^{1.03} \approx 2.80)$ increase in the odds of a Republican majority. Thus, Counties with higher depression are more likely to support GOP. However, this effect varies by state. (See *Random (state level) effects*)
- A 1 SD increase in proportion of white population i.e., by 15.88% (approx.) leads to 6.5 times $(e^{1.87} \approx 6.49)$ higher odds of a Republican majority. Thus, Counties with a higher proportion of white population are overwhelmingly Republican.
- A 1 SD increase in the log population i.e., by 1.51 units, equivalent to a 4.5x increase in population brings about a reduction of 65% $(e^{-1.05} \approx 0.35)$ in

5

the odds of a Republican majority. Thus, smaller, rural counties typically favor the Republican party.

- The standard deviations of predictors were calculated from the unscaled variables: 3.14% for depression, 15.88% for white population proportion, and 1.51 for log population

Random (state level) effects

Parameter	Estimate	95% CrI
SD(Intercept)	2.03	[1.55, 2.61]
SD(Depression Effect)	1.45	[0.96, 2.02]
Cor(Intercept, Depression Effect)	0.44	[0.04, 0.74]

- The standard deviation of the intercept (2.03) indicates that the baseline odds of a GOP majority vary significantly across states even if the other factors are not varied.
- The standard deviation of the slope of depression (1.45) also confirms that the effect of depression rates on GOP support is not uniform.
- The positive correlation between the state level intercept and the depression slopes suggest that in states where the baseline odds of a GOP majority is high, increase in depression rates is also accompanied by an increase in odds of GOP majority.
- The central figure (figure 3) illustrates how the effect of depression varies by state,
    - States like Idaho (Republican stronghold) and New Mexico show a strong positive effect i.e., in these states, higher depression states increase the odds of a Republican majority.
    - Whereas states like New Hampshire and Vermont (Democratic strongholds) show negative effects.
    - This variation suggests that regional factors influence the Depression-GOP relationship.
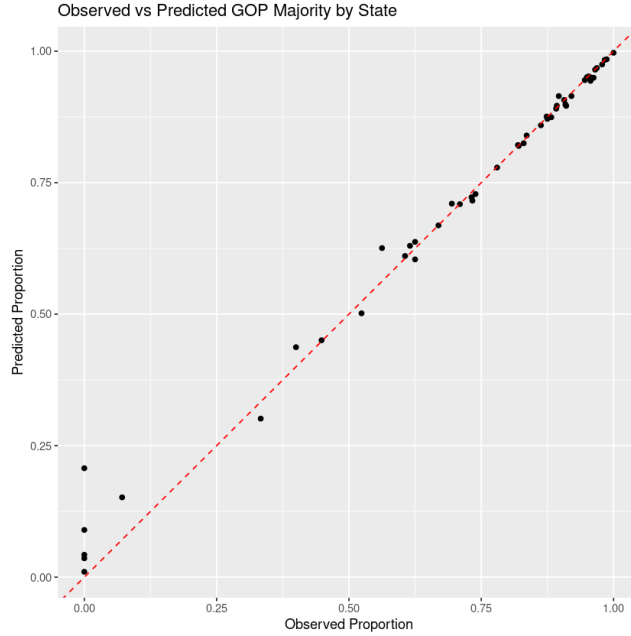
Predictive performance

Figure 5: Observed vs Predicted GOP Majority by State

Figure 5 illustrates the model's predictive alignment between observed and predicted values. The scatter plot compares the proportion of observed Republican majority counties in a state to model predictions. Most of the points lie on or around the line, indicating good performance, but some states deviate from the line reflecting the state level effects as seen in the central figure (figure 3). The Root mean squared error of the model's performance at predicting the state level proportions of GOP majorities is calculated to be 0.04, and the accuracy of the model at predicting if the county has a GOP majority is found to be 92%.

Conclusion

The analysis provides clear evidence of a positive association between depression rates and support for GOP in the 2024 election, albeit with strong state level variations. Counties with a higher proportion of white residents are more likely to vote for the Republican party, and so do counties with smaller populations. Both of these effects are also stronger than the association with depression rates, with race proportion exerting a powerful influence. Thus, states with smaller, homogeneous counties tend to have high support for GOP, the effect also amplified by higher depression rates. Conversely, states with larger, more

diverse counties tend to have lower GOP support, and in specific states, higher depression rates have a negative influence on GOP support.