YOUR TITLE

— Project Report —
Advanced Bayesian Data Analysis

Eldaleona Odole, Leonor Cunha, and Anarghya Murthy

February 27, 2025

TU Dortmund University

1 Introduction

notes to ourselves about things that should go in each section can be in red like this so we dont forget to delete them :)

Every two years the United States elects 435 officials to the House of Representatives. The 435 House seats are allocated roughly proportional to population with additional constraint that each state must have at least one seat in the House of Representatives. However, despite being roughly equivalent in population size, the characteristics of voters living in each district varies greatly by district.

Although technically a multiparty system, the U.S. is often called a two party system due to the domination of Democrats and Republicans at all levels of govenrment (cite). These parties dominate because political candidates are only required to get a plurality of votes, of which the two largest parties often reach. Additionaly would be third-party voters, will often vote for one of the two major parties so ensure their voice is heard, rather than using thier vote on a candidate who will likely not win (cite). Within our dataset, there are no districts represented by a third-party candidate and as such we often refer to the U.S. as being a two party system. The association between a voters demographics (gender, age, education etc.) and their propensity to vote for either a democratic or republican candidates is a topic of extensive study. However, little is known about the relationship between voting outcomes and the voters environment. In a very general sense, conventional wisdom says that cities tend to be more “blue”, meaning voters in large cities tend to vote for democrats. Our project is focused on characterizing this relationship more concretely.

To answer our question about the nature of the relationship between urbanizaiton and partisan voting outcomes, we choose to investigate the 2022 House Election. This election takes places during a non-presidential year, which may cause additonal effects, this election also takes place most recently after the 2020 Census, allowing us to use the most recently available demographic data.

Within this report we combine demographic data and urbanization data into a logistic regression model to predict the district voting outcomes of the 2022 House Election. The parameters

2 Dataset

Our dataset was made by combining three independent datasets related to the 2022 House election. The first dataset is the publically available urbanization dataset published by fivethirty eight from which we incorporate the variables urbanization index and (urban) grouping into our final dataset Holly Fuong, 2022. From the description of the dataset: ”The urbanization index is calculated as the natural logarithm of the average number of people living within a five-mile radius of every census tract in a given district, based on a weighted average of the population of each census tract. The population of a census tract is according to 2020 census data. This provides a numerical value for how urban or rural a district is. ” Holly Fuong, 2022. The urbanization dataset was put together by FiveThirtyEight as part of their analysis *The Republican Path To a House Majority Goes Through The Suburbs* which gave election predictions leading up to the 2022 U.S. Congressional Eleciton Skelly, n.d.

The second dataset used in our analysis the Election Results Dataset from FiveThirtyEight Mehta, n.d. It is a continuously updated repository of United States Govenor, Congressional and Presidential elections. As this dataset includes all elections going back to 1998, we only used a subset of the data relevant to the 2022 House Election. From this dataset we used the party, state, and winner variables.

The third data used in our analysis is a subset of the 2022 American Commuity Survey Data. The American Community Survey is a yearly survey collecting information about the occupations, education attainment, income and other demographic information carried out by the United States Census Bureau. The United States Census Bureau provides an online tool to access its extensive survey database, which can then be filtered and refined for further analysis. For our analysis we used the following variables for each House district;

2.1 Data Cleaning

In the initial cleaning we wanted all the variable to be on roughly the same scale to aid in convergence times. In order to do that we roughly scaled median income and total population by dividing total population by one million and dividing median income by one hundred thousand. This brought each of these to roughly the same scale as the other variables that are in the range of zero to one as they are percentages.

3 Models

The Winning party in each congressional district race ($y_{i,j,k}$ for district i, state j, region k) can be modeled as the outcome of Bernoulli trial, since this is a binary variable: NOTE ON HOW WE KNOW IT IS NOT A 2 PARTY SYSTEM EXCEPT THAT OOPS IT IS

$$y_{i,j,k} \sim Ber\left(\pi_{j,k} = logit^{-1}(\theta_{j,k})\right) \tag{1}$$

with probability of a Democrat win $\pi_{j,k}$ modeled as the inverse logit transform of $\theta_{j,k}$, a linear combination of our covariates. The inverse logit function converts real numbers into quantities between 0 and 1, and is therefore a standard way to model probabilities cite.

We tested four different models for $\theta_{j,k}$, which include different covariates in addition to out variable of interest (Urban index) and incorporate our data's hierarchical structure in different ways. Therefore, all four are Multilevel Bayesian (Logistic) Models, which require particular assumptions: CITE first, that a logistic regression accurately represents the relationship between the log-odds of a Democrat win and the explanatory variables, that is, $\theta_{j,k}$ and our covariates are linearly related; second, interchangeability, meaning that each district is exchangeable within each state and each state is exchangeable within each region; and third, that the value of urban index (and other covariates) in a district has a different effect depending on the state/region it belongs to.

The logistic relationship is a common assumption in the literature (cite). We can assume interchangeability because for complicated historical reasons certain regiions of the united states are more similar to eachother than others. For example the Southern United states tends to be more religious and religous people tend to vote more conservatively, as a result the parameter associated with region would likely be smaller or more negative as compared to other regions. The idea behind the differing effect strength of urban index values per region is that a city in a rural area will likely have stronger signal than an city among a bunch of other cities. [rewrite some of this paragraph]

Model 1

Model 1 is our most extensive model. Here we used urban index and 4 additional covariates plus an intercept to explain $\theta_{j,k}$. Both the state and region hierarchies were included, but on different covariates. Urban index and median income effects vary by state, while the slope of percentage of bachelor's degrees varies by region. The intercept and the slopes of percentage of women and percentage of retirees were considered to have the same effect for all districts, hence were modeled non-hierarchically.

$$\begin{aligned}\theta_{j,k} =&\beta_0 + \beta_{women} \cdot \text{Pct\_Women} + \beta_{uncent\,urbindex,j} \cdot \text{Urban\_Index} \\ &+ \beta_{uncent\,bsc,k} \cdot \text{Pct\_Bach.} \\ &+ \beta_{uncent\,inc,j} \cdot \text{Median\_Income} + \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{2}$$

Equation 2 describes our model conceptually. In order to have a specification more compatible with R syntax so we can fit our model with brms, we reformulate the model as Equation 3, with the group-level (state or region) centered around zero, following cite brms book

$$\begin{aligned}\theta_{j,k} =&\beta_0 + \beta_{women} \cdot \text{Pct\_Women} \\ &+ \beta_{urbindex} \cdot \text{Urban\_Index} + \beta_{urbindex,j} \cdot \text{Urban\_Index} \\ &+ \beta_{bsc} \cdot \text{Pct\_Bachelor's} + \beta_{bsc,k} \cdot \text{Pct\_Bach.} + \beta_{inc} \cdot \text{Median\_Income} \\ &+ \beta_{inc,k} \cdot \text{Median\_Income} + \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{3}$$

Model 2

For Model 2 we significantly reduced the number of variables. We kept only our variable of interest, urban index, and percentage of retirees [why?????] Here, the geographical hierarchy was incorporated through a *nested hierarchy* of districts within states within regions. So, this model assumes that the effect of urbanindex ($\beta_{urb,j:k}$) depends on state j and region k through a prior with mean parameter $\beta_{urb,k}$, which in turn varies by region and depends on hyper-mean β_{urb} (which has its own prior, with hyper-hyper-parameters).

Equation 4 specifies the model, with the centered around zero formulation.

$$\begin{aligned}\theta_{j,k} =& \beta_0 + \beta_{urb} \cdot \text{Urban\_Index} + \beta_{urb,k} \cdot \text{Urban\_Index} \\ &+ \beta_{urb,j:k} \cdot \text{Urban\_Index} + \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{4}$$

Model 3

$$\begin{aligned}\theta_{j} =& \beta_0 + \beta_{urb} \cdot \text{Urban\_Index} + \beta_{urb,j} \cdot \text{Urban\_Index} \\ &+ \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{5}$$

Model 4

$$\begin{aligned}\theta_{k} =& \beta_0 + \beta_{urb} \cdot \text{Urban\_Index} + \beta_{urb,k} \cdot \text{Urban\_Index} \\ &+ \beta_{ret} \cdot \text{Pct\_Retirees}\end{aligned} \tag{6}$$

4 Priors

	Model 1	Model 2	Model 3	Model 4
Intercept	$\beta_0 \sim N(0, 10)$	$\beta_0 \sim N(0, 10)$	$\beta_0 \sim N(0, 10)$	$\beta_0 \sim N(0, 10)$
Urban Index	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,j} \sim N(0, \sigma_{urb,j})$, $\sigma_{urb,j} \sim Gamma(2, 5)$	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,k} \sim N(0, \sigma_k)$, $\sigma_k \sim Halfcauchy(10)$ $\beta_{urb,j:k} \sim N(0, \sigma_{j:k})$, $\sigma_{j:k} \sim Halfcauchy(10)$	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,j} \sim N(0, \sigma_j)$, $\sigma_j \sim Halfcauchy(10)$	$\beta_{urb} \sim N(0, 1)$ $\beta_{urb,k} \sim N(0, \sigma_k)$, $\sigma_k \sim Halfcauchy(10)$
Pct.retirees	$\beta_{ret} \sim t(1, -2, 1)$	$\beta_{ret} \sim t(1, -2, 1)$	$\beta_{ret} \sim t(1, -2, 1)$	$\beta_{ret} \sim t(1, -2, 1)$
pct.women	$\beta_{urb} \sim N(0, 1)$			
pct bsc	$\beta_{bsc} \sim t(1, 0, 1)$ $\beta_{bsc,k} \sim N(0, \sigma_{bsc,k})$, $\sigma_{bsc,k} \sim Halfnormal(0, 1)$			
median income	$\beta_{inc} \sim N(0, 1)$ $\beta_{inc,j} \sim N(0, \sigma_{inc,j})$, $\sigma_{inc,j} \sim Halfnormal(0, 1)$			

Table 1: Prior summary table

For the intercept β_0 in Model 1 we set a Normal prior centered at zero with a large standard deviation. This represents a weakly informative prior, as we had no strong beliefs about the intercept value, nor does it have any straightforward interpretation in our model: it theoretically represents the (logit of the) probability of a Democrat win in a district with no urbanization at all, a median income of zero dollars, and 0% of women, retirees and citizens with a bachelor's degree in the population; such a district is obviously nonexistent.

In Model 1, neither Pct.Women nor Pct.Retirees are modeled hierarchically.

The percentage of women is roughly the same in every district, so we do not expect this covariate to have a strong effect on the probability of either party winning, i.e, we expect β_{women} to be close to zero. So, we set a prior for this slope which is centered around zero and has little variability: a standard normal prior.

The Percentage of retirees in each district negatively correlates with the probability of a Democrat winning, but we do not know how strong this effect ought to be. Therefore, for the prior on β_{ret} we chose a distribution centered around a negative number, and with relatively heavy tails, reflecting our uncertainty.

The effects of urban index, percentage of bachelor degrees, and median income are all parameterized in 2 hierarchical levels: an average slope across all districts, and a varying slope by group (State or Region), $\beta_{covariate,j}$ or $\beta_{covariate,k}$, which follows a Normal distribution centered at zero with standard deviation modeled at group level (by a hyperprior).

For the population-level component of the urbanindex slope we opted for a standard normal prior. We chose not to make assumptions on the sign of the effect of this variable, as it is this variable that we are interested in studying. So, we set comparatively less-informative priors on the parameters representing the effect of the index.

As we expected the population-level effects of both Median Income and Pct Bscs to be positive in some cases and negative in others, we picked symmetric priors for both β_{bsc} and β_{inc}. We are, however, less sure about the average null effect of the Percentage of Bachelor degrees, so for this slope parameter we opted for a prior with 'fatter tails', the standard Cauchy distribution rather than the Normal one.

All group-level (zero-centered) priors are Normal, by brms specification is there a reason???

For the hyperparameters, we chose a half standard normal prior for both the standard deviations of $\beta_{bsc,k}$ and $\beta_{inc,j}$. This is a narrow distribution, with most values falling between 0 and 1, as we expect to see weak effects for these covariates, and thus small standard deviations (and positive, as any SD is by definition). For the standard deviation of $\beta_{urb,j}$, on the contrary, we opted for a less informative prior. Again, we do not want to make such strong assumptions about the effect of our variable of interest, hence we "allow the estimates to fluctuate more".

5 Code

6 Results

This section is not on the instructions but is probably the easiest way to talk about the results we got

7 Convergence Diagnostics

8 Model Comparison

Our four models were built based on somewhat different assumptions about the structure of our data, and all produced slightly different results. We need to know which of these is *better*, that is, which results are more trustworthy and allow us to answer our original research question. To this end, we measured and compared our models' predictive performance first by looking at absolute predictive performance, then at relative and finally at the Leave-One-Out statistics to compare out-of-sample Predictive Performance.

To measure absolute predictive performance we used the Root Mean of Squared Error (RMSE) (FORMULA???). This measure works in a similar way to the R squared statistic that is commonly used to assess the fit of linear models, by evaluating differences between observations and model predictions, but RMSE retains the scale of the response variable, meaning it has a direct interpretation in the context of our problem. It takes into account the uncertainty of the posterior distribution by...

[plots of RMSE draws, overlap for comparison]

[interpretation]

To assess relative predictive performance, we looked at log-likelihood scores, that is, the average of posterior draws' log-likelihoods for each observation (FORMULA???). This is a relative predictive performance measure in the sense that it does not tell us anything about the model's predictive performance alone, we need to compare it between different models to establish which is better.

[plot of ll scores or likelihood differences?, overlap]

[interpretation]

In-sample predictive performance measures evaluate only model predictions for the same observations which were used to fit the model in the first place, therefore they tend to favor more complex models. (In our case, the bigger mmodel (Model 1) was indeed the preferred one using both RMSE and LL scores.????????????????') Because we are comparing models with different degrees of complexity, it is essential to check also out-of-sample predictive performance metrics. These metrics are computed by splitting the dataset into training data and test data, fitting the model on the former and assessing the likelihood (ELPD) of the observations in the latter, given the model estimates with the training set. ELPD FORMULA??????????????

The way we choose to split the data into training/test sets naturally impacts the ELPD. So, we rely on cross-validation: we do multiple different splits and average over the results. Our chosen method was Leave-One-Out cross-validation, which in theory performs as many splits as observations in the dataset, each time leaving one "out" as the test data. In practice, a different posterior is not actually computed this many times, but rather an estimate from the full model posterior using importance sampling (PSIS).

[LOO statistics table]

According to the LOO statistics, model 1 is the preferred one.

[pareto k estimates issue and momnet matching?]

9 Prior Sensitivity Analysis

another table with priors here, maybe not all because thats a lot

One of the most important parts of Bayesian Data Analysis is setting the prior distributions. The choice of priors could greatly affect the final results in a model.. (cite prior sensitivity guys) So, we conducted a prior sensitivity analysis, by refitting our model(s) using alternative priors (which also fit our model assumptions) and assessing the impact in our results.

[BRIEFLY explain new priors, graphs comparing them]

10 Limitations and Improvements

11 Conclusion

11.1 Reflection on own learnings

please lets call this subsection something else, this sounds so childish

References

Holly Fuong, G. S. (2022). District urbanization index 2022. https://github.com/fivethirtyeight/data/tree/master/district-urbanization-index-2022

Mehta, D. (n.d.). Election results. https://github.com/fivethirtyeight/election-results/blob/main/README.md

Skelly, G. (n.d.). *The republican path to a house majority goes through the suburbs*. https://fivethirtyeight.com/features/the-republican-path-to-a-house-majority-goes-through-the-suburbs/