Investigating Voter Turnout in Indian Elections

Stat 5293

Jared Lieberman

The goal of our project is to use quantitative methods we have learned in this class in an applied social science setting. We are exploring potential factors of why voter turnout in India tends to be higher in local and state-level elections vs. national elections. Among democracies with constant and relatively stable elections, this trend in India is anomalous compared to elections in the United States and Europe. We are interested in differences at each level, and multilevel modeling is particularly useful for our case because of the large regional and state differences in India. Some potential factors we are exploring include time variation between elections, urbanization, and how competitive elections are.

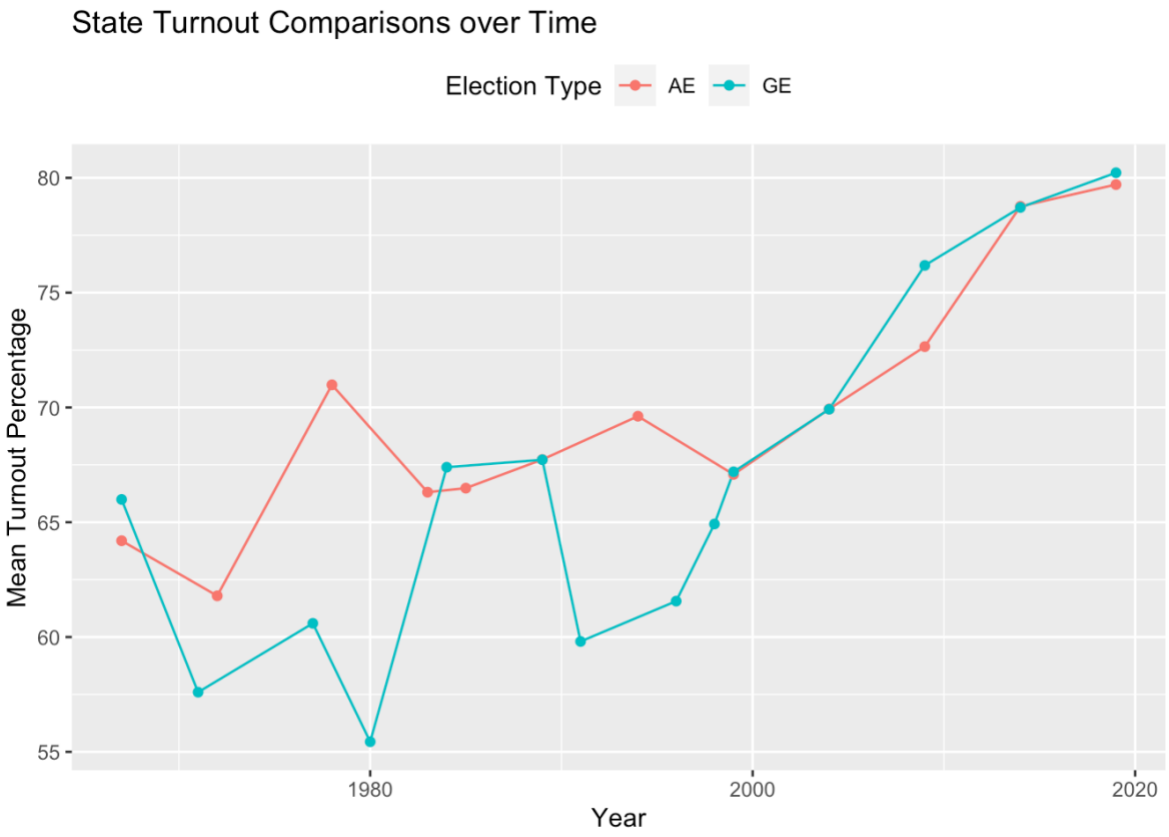
Due to the Indian voting system's unique set up, data is presented in a three-level structure, with the assembly constituency (lowest level) contained within the parliament constituency, which is then contained within the state (highest level). Our project aims to structure such a three-level model which can accurately utilize and predict the turnout difference at different levels. Our dataset contains voting data from 1967 - 2021 over 28 states in India with more than 500 parliament constituencies. India has different periods of time in which constituencies are redrawn, called delimitation. In order to simplify matching constituencies over time, we used data from the third delimitation from 1974 - 2007. We ultimately created a large dataset with 25,889 rows, each of which corresponds to an assembly constituency election (AE), which are matched to general elections (GE) at the parliamentary level. For every election, we have data on the number of votes and electors, the number of candidates, ENOP (Effective Number of Parties), demographic information on the constituency, demographic information on the winner (i.e. incumbent), the time difference between matched elections, and urbanization data. Our project also has the technical goal of constructing a computationally efficient model to present our unique data structure and answer our questions about the turnout difference.

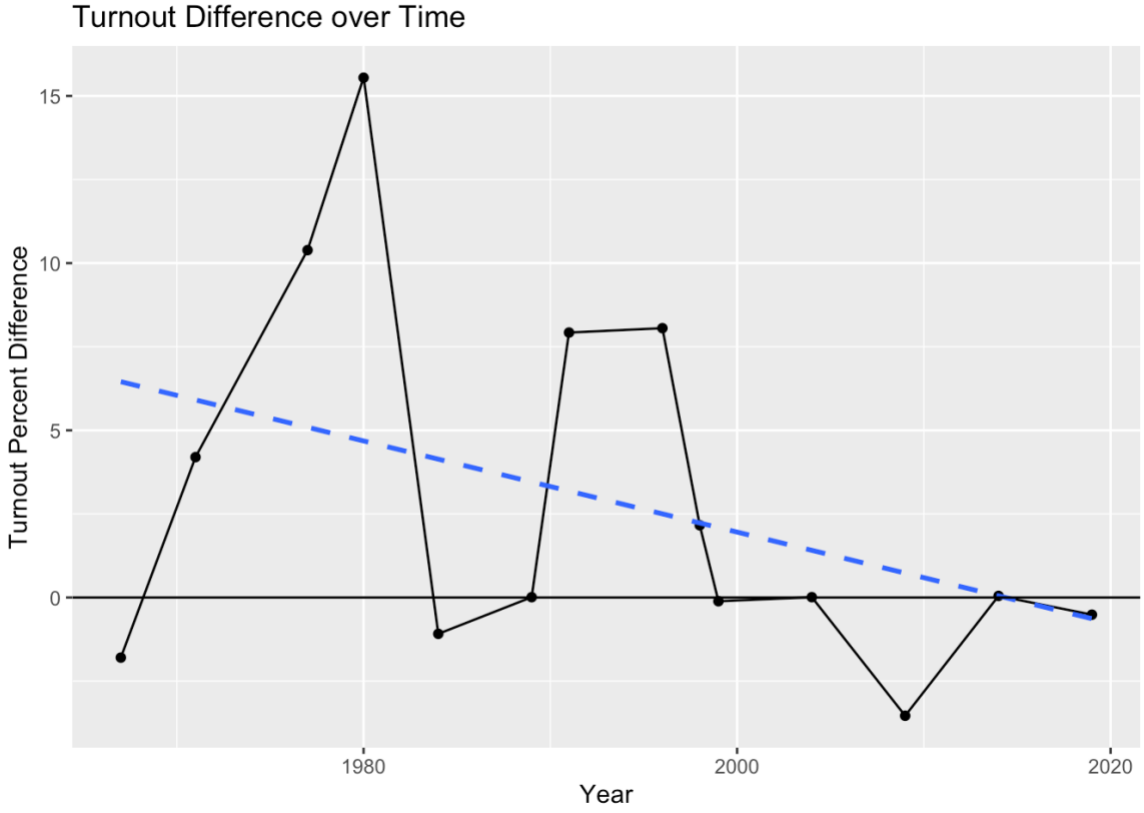
The data were initially collected from the election dataset at the Trivedi Centre for Political Data (TCPD), a research lab that was housed in the Political Science Department at Ashoka University in Sonepat, India. This data includes information on every election at the Parliamentary and Assembly levels, similar to national- and state-level elections, respectively, in the United States. The raw data that TCPD used is from the Election Commission of India (ECI). For the number of parliamentary seats in each

state, we use a dataset from the ECI. We use a dataset on urbanization that was published by Kanchan Chandra and Alan Potter, affiliated at NYU. We are currently in the process of merging these datasets in order to prepare a nested format between states, parliamentary elections, and assembly elections.

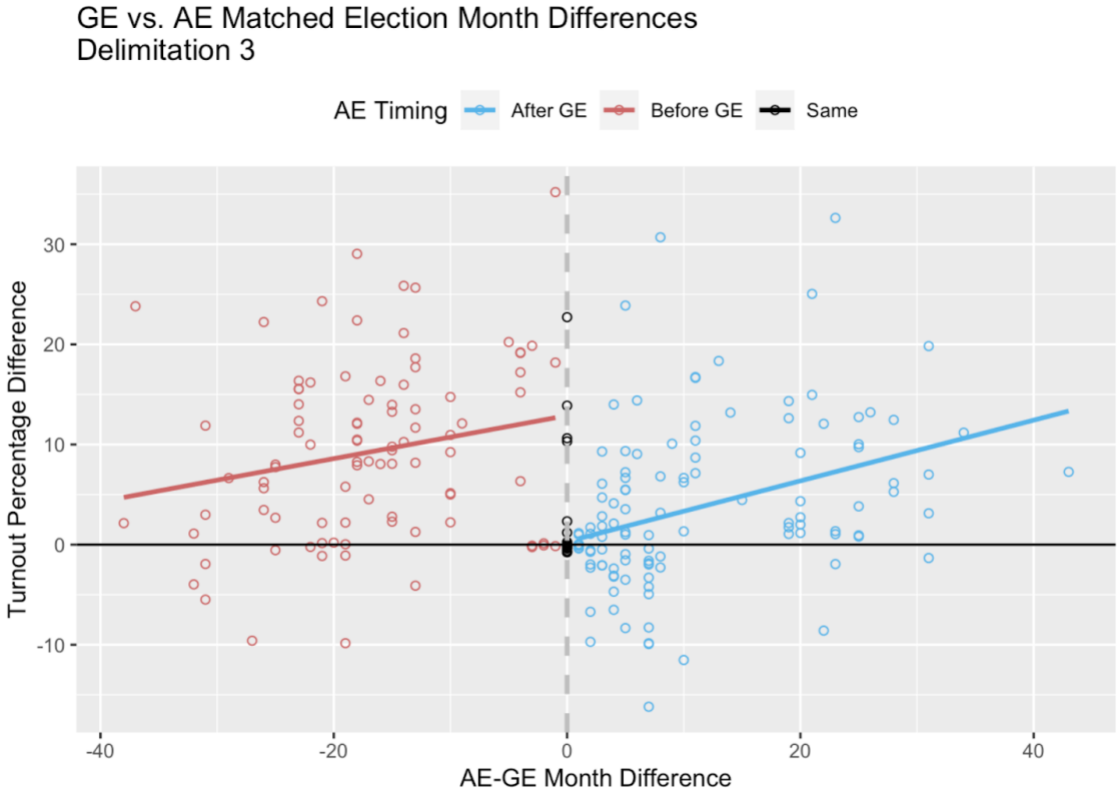
The process of cleaning and preparing the data for the models has been particularly tedious. Because every assembly constituency is contained within a parliamentary constituency, we needed to match elections between the two in order to get the turnout differences. The methodology we used for the matching was to find the closest election between AE and GE by month and year. Because there is not a one-to-one relationship between the elections, there are sometimes repeated assembly elections. In order to account for these repetitions, we created two datasets: one full dataset with repeated GE matches and one without repeats. We removed the repeated rows that were not the closest election match, time-wise, of the repeated AE-GE matches. From the matches, we were then able to calculate the turnout differences, which was simply the value of the AE turnout minus the value of the GE turnout. Thus, if the difference value is positive, then the AE turnout is higher. As a note, the urbanization data contains density values at the parliamentary constituency level. In order to more accurately capture the process of urbanization over time, we created a lagged value to get the percent change in density.

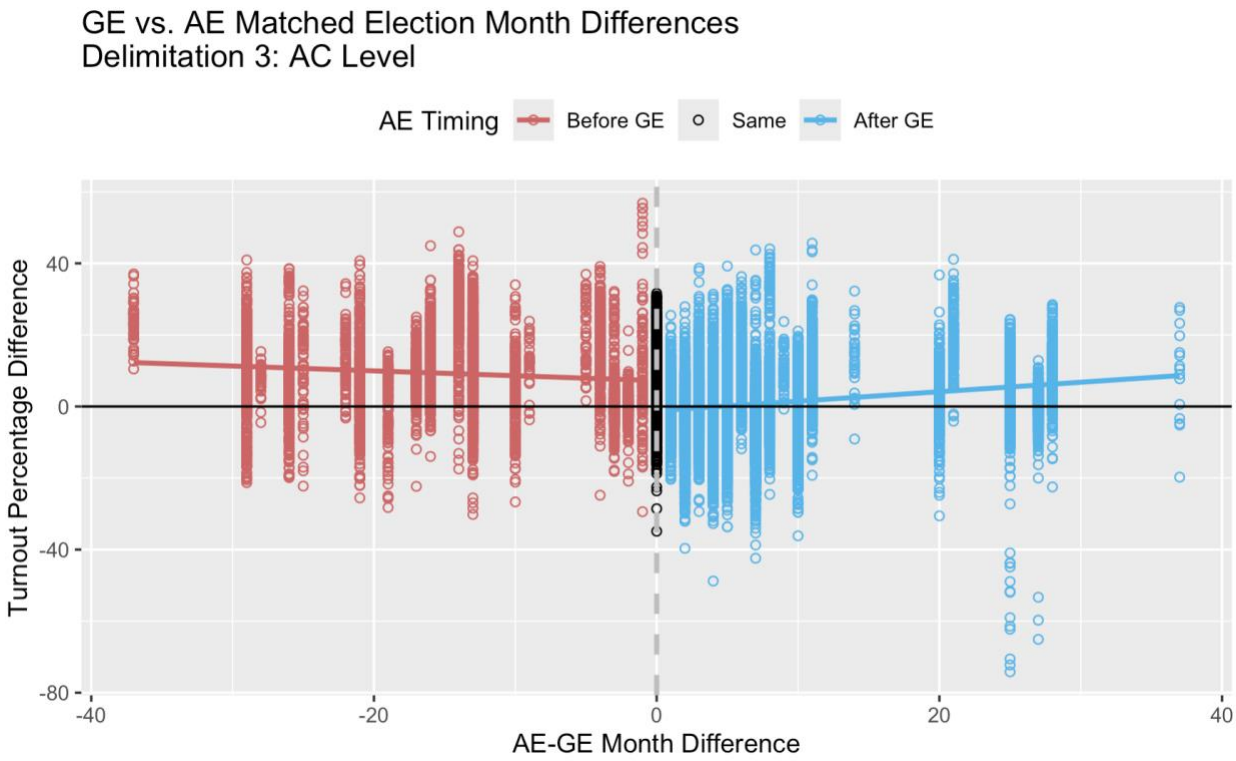
We conducted exploratory analysis at the state level in order to have an initial understanding of differences between states and over time. For example, we have two graphs below for the state Andhra Pradesh. We created similar graphs for every state. The graphs use data from 1967 - 2021, and we can see the mean turnout percentages over time split by AE and GE. This is illuminating because we can see there is an increase in voter turnout over time, regardless of election type. In the second graph, we can see the turnout difference over time for Andhra Pradesh. This shows us that the difference is decreasing, perhaps in part to the nationalization of politics in India.


State Turnout Comparisons over Time
Election Type
AE
GE
Mean Turnout Percentage
80
75
70
65
60
55
1980
2000
2020
Year


Turnout Difference over Time
Turnout Percent Difference
15
10
5
0
1980
2000
2020
Year

As noted, we incorporate time differences into the analysis, in particular the month difference between general and assembly elections. Below are discontinuity regression graphs of the turnout differences across time variation. The first graph contains election matches aggregated at the Parliamentary Constituency level. The second contains election non-aggregated matches at the Assembly Constituency level. The outliers are from smaller states and particular years, and thus it would be interesting as a next step to remove them to see if the regression lines change. In particular, there are extreme values in which the General Elections were far greater than the matched Assembly Elections. Additionally, we could have a time period cut-off when the matched elections are too far apart to measure comparisons.





We then like to fit a varying slope and varying intercept model for our dataset to accurately predict the turnout difference between assembly election and general election in India. However, we are cautious with the fact that with a multilevel model of this scale, it can be computationally demanding for the large number of states and constituencies of different levels, but also for the multiple folds of computation by the multilevel fitting, especially by Bayesian models. Therefore, constructing a model in the most efficient way is of a priority.

In the first iteration of data preparation, we had a dataset of two levels, state and parliament constituency. We considered using lmer and stan_glmer to fit a varying slope and intercept model. Due to concerns that our potential model takes too much computational power, we have decided against using all variables available as predictors for our outcome due to both validity and computational concerns. We have, therefore, selected only four predictor variables: month difference (the number of months apart between the general election and the closest assembly election), urbanization density (an urbanization index), electors (population in parliament constituency) and the general election year. We first considered a model as such:

```
Turnout.Difference ~ Year.GE + (1 | State_Name)
```

This model takes the general election year as the single predictor variable and a varying intercept by states. From this model, we were able to conclude that the general election year has a positive effect for the turnout rate difference, specifically we expect for the increase of election year to have around 0.24% increase in the turnout rate difference. Next we included all four variables in our second model:

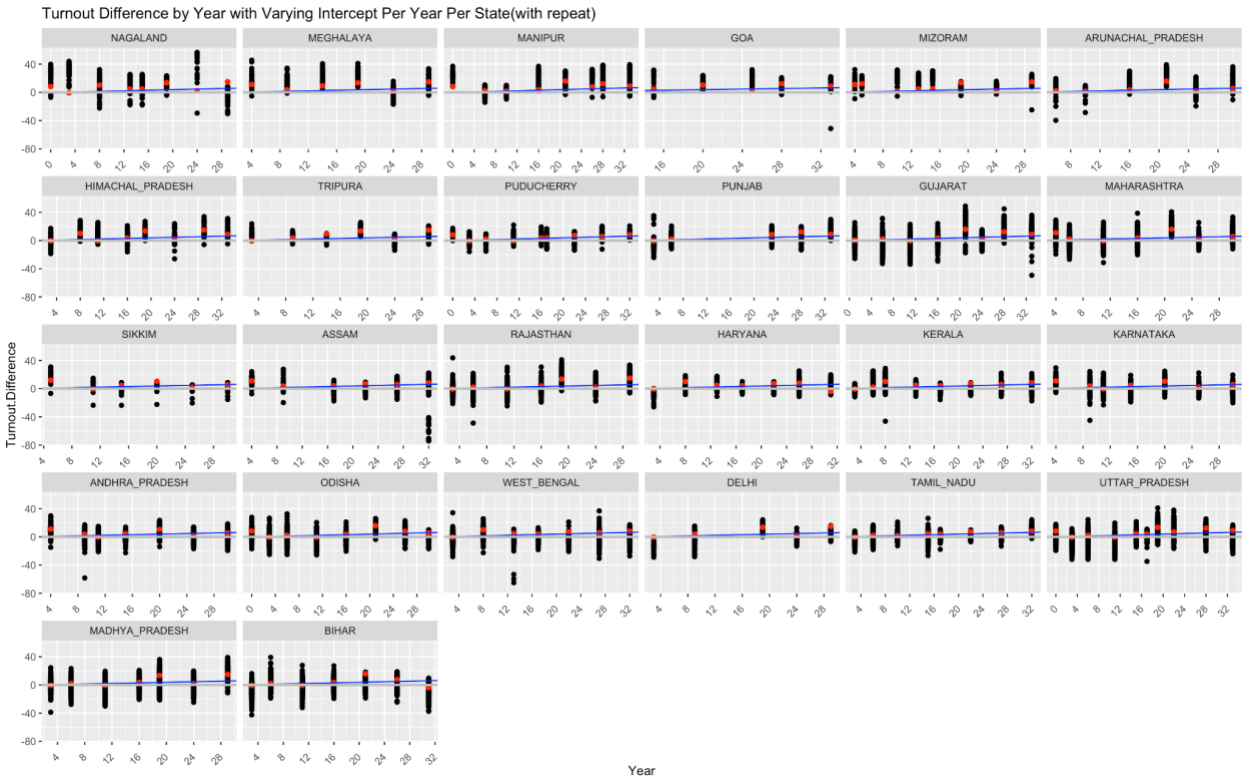
```
Turnout.Difference ~ Month.Difference + Density + Electors + Year.GE
+ (1 + Month.Difference | State_Name)
```

This model includes the month difference, urbanization density, electors and general election year as the fixed effect part of the model and consider month difference as the random effect part of the model. From this model, we found that the month difference has a negative effect on the turnout rate difference and generally election year has a positive one. Controlling for other variables, we would expect a year passed and a month difference in the two elections to have a .14% increase and -0.1% in turnout difference respectively.

In the newest round of data analysis, we focused on two specific types of models and their results, and both models will take in both datasets with and without repeat. The first of such is:

```
Turnout.Difference ~ Year + (1 | State_Name) + (1 | Year), data =
with repeat data
```

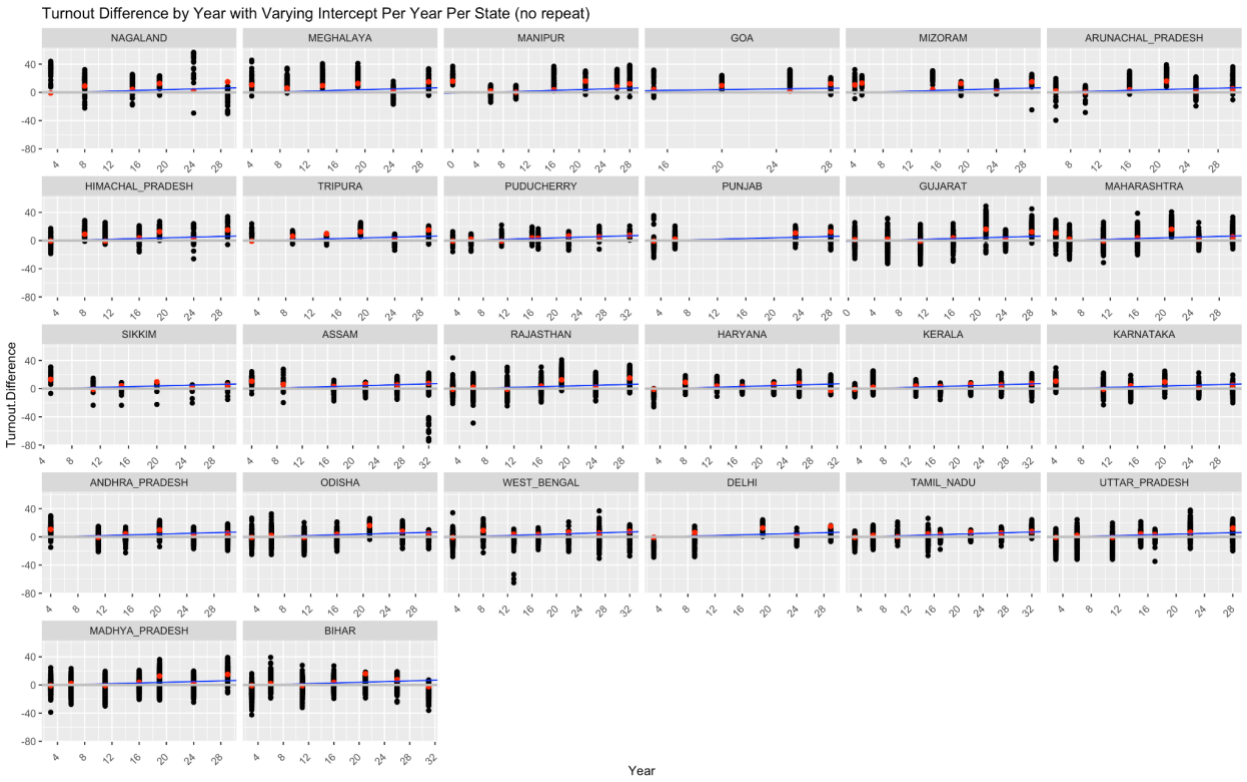
This model takes the assembly election year as the single predictor variable and a varying intercept by states and by year. We have found that the year is not a significant predictor for the turnout rate difference. The variation between the election years appears to be similar with the variation between the states. We unfold the model plotting the turnout difference vs. year in each of the states. The red dots and blue line represent the turnout rate difference prediction adjusted for yearly and state intercepts and the complete pooling regression model.



Similarly we also fitted the same model on data with no repeat:

`Turnout.Difference ~ Year + (1 | State_Name) + (1 | Year), data = no repeat data`

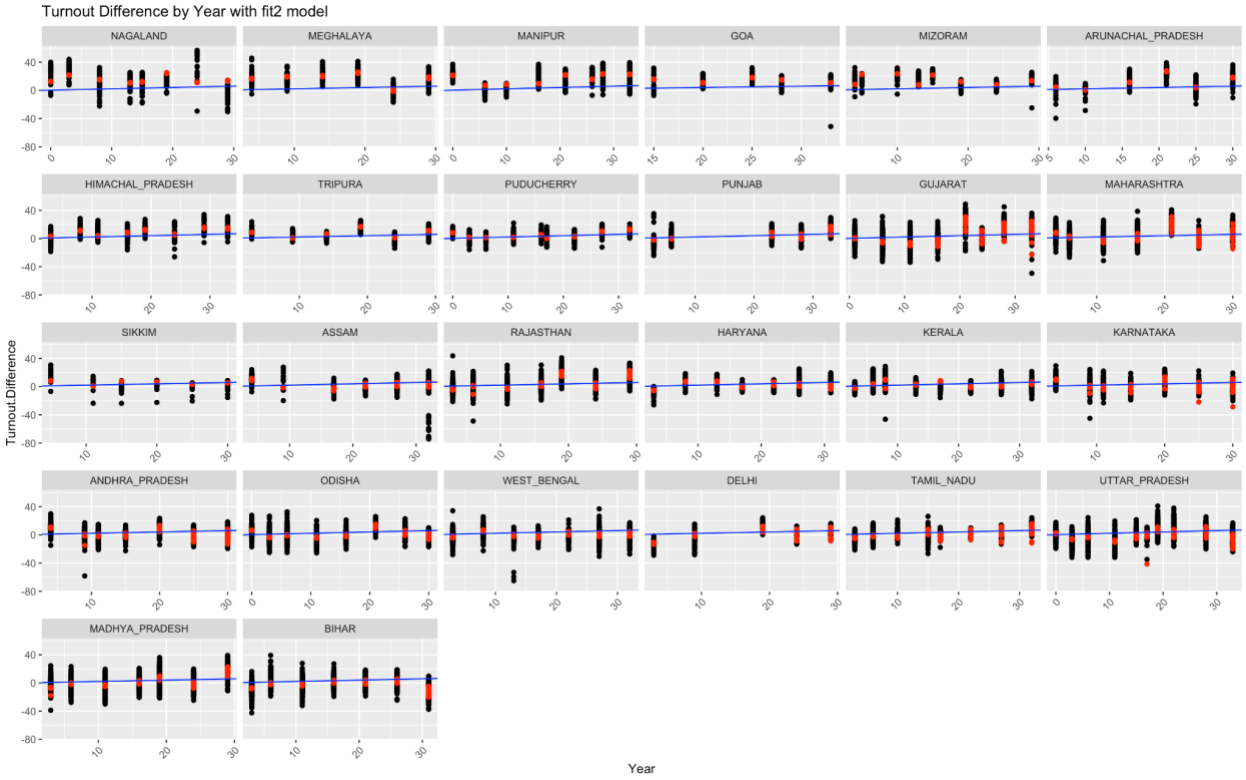
On this dataset, we have less number of data available than the repeated dataset, which further utilizes the multilevel partial pooling model's advantage of overcoming a smaller dataset. We have found that the year remained to be an insignificant predictor of turnout rate difference. We also observe a larger variation for random effect between election years, likely due to the smaller sample size. Similarly we unfold the model by plotting the data and estimates by state. The red dots and blue line similarly represent the turnout rate difference prediction adjusted for yearly and state intercepts and the complete pooling regression model respectively.



We then further explored models that included more predictors

```
Turnout.Difference ~ Month.Difference.abs + AE.Before.GE +
Electors.AE + Electors.GE + Year + No.PC + (1 + Month.Difference.abs +
AE.Before.GE | State_Name) +                    (1 | Year), data =
data_repeat
```

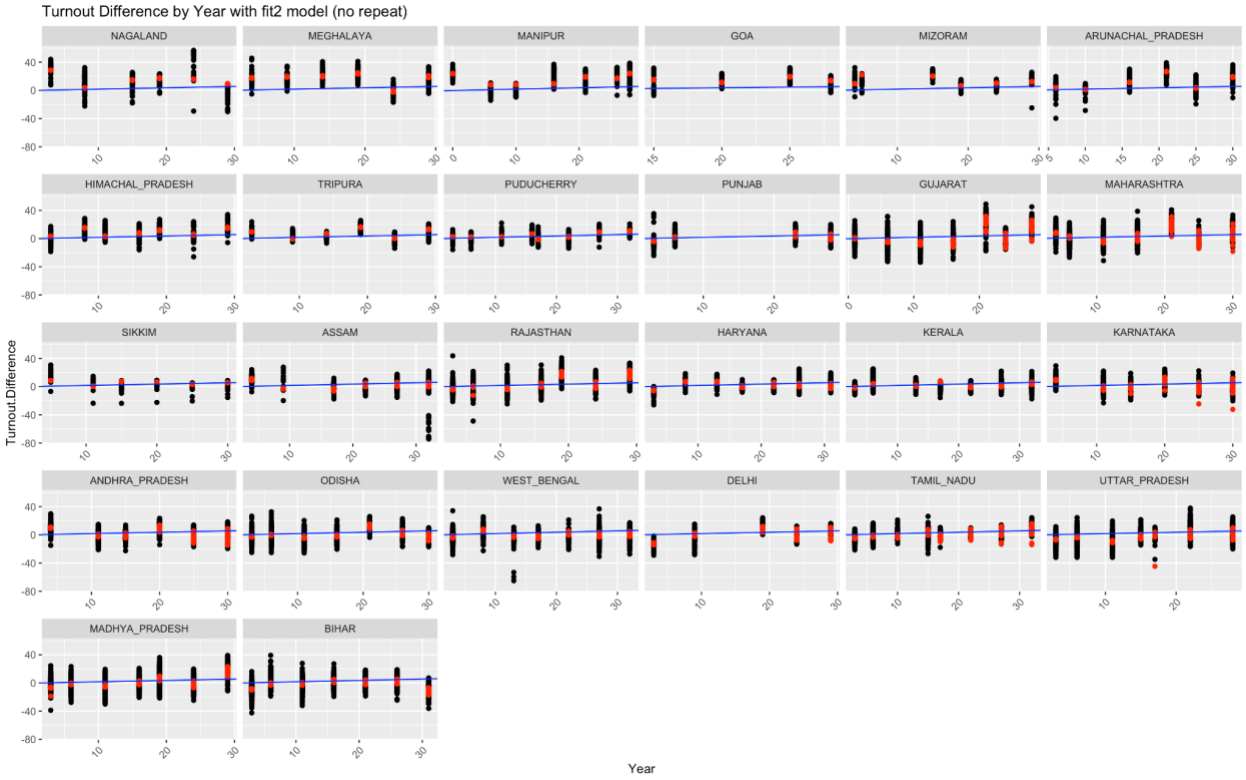
In this model, we take more predictors: the number of months apart between the two matched elections, an indicator variable for if the assembly election happens before the general election, a lagged density for the urbanization change over time, the electors for both the general and assembly elections, the year of assembly election as well as the number of constituencies within each state. We also include in random effect: a varying intercept for each state and each year, and varying month difference and order variables by state as well. From this model, we see that the absolute number of months apart between the two elections and the number of parliament constituencies are the only significant predictors of the turnout rate difference. On the other hand, we have observed that the variability for different years is quite large.The effectiveness in whether AE is before GE predicting the turnout rate difference is noisy between states as well. We have also used our mixed-effects model to predict values which are plotted in red dots, and the completely pooled model is graphed in blue.



The same model was also fitted for the no repeat data:

```
Turnout.Difference ~ Month.Difference.abs + AE.Before.GE +
Electors.AE + Electors.GE + Year + No.PC + (1 + Month.Difference.abs +
AE.Before.GE | State_Name) +                    (1 | Year), data =
data_no_repeat
```

When the same model fitted on the no-repeat dataset, we found that the list of significant predictors expanded from the absolute number of months apart from the two elections and the number of parliamentary constituencies to include whether or not the assembly election is before the general election. However we still find the indicator variable to have a large variability between states similar to our last model fit on the repeated dataset.



Regarding additional analysis, what we could have done differently, and what we did wrong, we believe there are still questions regarding the methodology of comparing assembly and general elections in order to understand turnout rate differences. For example, Professor Gelman mentioned in our second presentation that we could create a “long” dataset, as opposed to our current “wide” dataset, in order to not have to work through some of the matching problems and redundancies. Additionally, we could add interaction terms about what type of election (AE vs. GE) and election timing (AE before vs. after GE) on the other predictors. As noted, because of the diverse nature of the Indian electorate and the critical role that multilevel modeling plays accounting for vast differences between states and constituencies, we would not want to lose the hierarchical structure of the data.