Statistics Comprehensive Project Report
Election Analysis with Special Reference to Demography

Group 6
Anvit Garg - BS1836
Prasun De - BS1826
Rajdeep Brahma - BS1821

October 7, 2022

Abstract

We analyze the previous three West Bengal Legislative Assembly elections. We start with trend analysis for each constituency and district. Next, we check if there is 'Demographic Divide' among the various demographic groups based on the three WBLA Elections. In Statistical terms, Say C is a class of voters in which the voters are assumed to be iid (henceforth referred to as blocks) and A is a political party/alliance. We wish to find if there is any significant relation between C and votes to party A. We also find if there are any districts which behave abnormally as compared to the state via Mahalanobis distance.

1 Data Collected

The following data was collected

1. Demographic data for West Bengal based on Census 2011 [2]
2. West Bengal Legislative Assembly data for 2011 [3]
3. West Bengal Legislative Assembly data for 2016 [4]
4. West Bengal Legislative Assembly data for 2021 [5] [1]

Pre Processing

Minimal amount of pre-processing was required for 1, mainly to estimate constituency wise data when it was missing. Extensive Pre Processing was required for 3. since the data was in PDF format and had to be converted and formatted manually. No pre-processing was necessary for 2. We had to write a program to scrape data for 2021 election since a Statistical Report had not been published yet.

2 Methodology

We propose the following methods that will help us reach a conclusion

1. Trend analysis by categorizing different constituencies and districts into 4 classes based on TMC votes.
2. Say $P(x \text{ votes for } A | x \in C) = p_1$ and $P(x \text{ votes for } A | x \notin C) = p_2$. Then we will estimate p_1 and p_2 via simple linear regression.
3. We try to combine these estimate into meaningful probabilities via restricted multivariate regression.
4. Calculating Mahalanobis distance for outlier detection.

3 Variables

All of the variable are in proportion of the total population of the constituency

- Ub : Urban
- Hn : Hindu
- M : Males
- Work : Working people
- Lit : Literate people

We also had interaction between Ub and the other 4 variable and between Hn and other three variables. For example, UbLit means proportion of people who are literate and live in urban areas. Since census blocks and constituency aren't same everywhere, some of the constituency demographics had to be estimated by a superset or a subset.

We also look at the votes polled by the three major coalitions LF , AITC and NDA and corresponding to them we have the following variables :

- LF11: proportion of votes obtained by LF in 2011
- LF16: proportion of votes obtained by LF in 2016
- AITC11: proportion of votes obtained by AITC in 2011
- AITC16: proportion of votes obtained by AITC in 2016
- AITC21: proportion of votes obtained by AITC in 2021
- NDA21: proportion of votes obtained by NDA in 2021

Further, we also have number of electors and number of votes casted for all three years.

4 Preliminary Observations

We made the correlation matrix for the predictor variables

	Ub	Hn	M	Lit	Work	UbHn	UbM	UbLit	UbWork	HnM	LitHn	WorkHn
Ub	1	0.268498	0.211092	0.555631	-0.33164	0.969896	0.999792	0.993285	0.993813	0.270845	0.447214	0.069473
Hn	0.268498	1	-0.00699	0.577564	0.26057	0.382847	0.267668	0.296006	0.251025	0.999158	0.931355	0.929131
M	0.211092	-0.00699	1	0.08882	-0.27925	0.187852	0.226924	0.204499	0.202474	0.030479	0.048476	-0.09157
Lit	0.555631	0.577564	0.08882	1	-0.10753	0.589393	0.554537	0.604024	0.55162	0.576403	0.813113	0.435839
Work	-0.33164	0.26057	-0.27925	-0.10753	1	-0.33075	-0.33173	-0.32412	-0.28644	0.25279	0.107016	0.579055
UbHn	0.969896	0.382847	0.187852	0.589393	-0.33075	1	0.969012	0.976357	0.957009	0.384157	0.545691	0.170105
UbM	0.999792	0.267668	0.226924	0.554537	-0.33173	0.969012	1	0.992917	0.993473	0.270667	0.445873	0.069014
UbLit	0.993285	0.296006	0.204499	0.604024	-0.32412	0.976357	0.992917	1	0.988031	0.297468	0.488644	0.098505
UbWork	0.993813	0.251025	0.202474	0.55162	-0.28644	0.957009	0.993473	0.988031	1	0.253149	0.434941	0.069556
HnM	0.270845	0.999158	0.030479	0.576403	0.25279	0.384157	0.270667	0.297468	0.253149	1	0.930808	0.925718
LitHn	0.447214	0.931355	0.048476	0.813113	0.107016	0.545691	0.445873	0.488644	0.434941	0.930808	1	0.812194
WorkHn	0.069473	0.929131	-0.09157	0.435839	0.579055	0.170105	0.069014	0.098505	0.069556	0.925718	0.812194	1

We see that the interaction effects are very highly correlated with the main factors. Hence, there is no point in using them to fit the model. The predictor "Males" was not used because the proportion of Males all the constituencies is almost the same. It only ranges from 0.5 to 0.53. Further analysis showed that the variable "Work" was not useful as well.

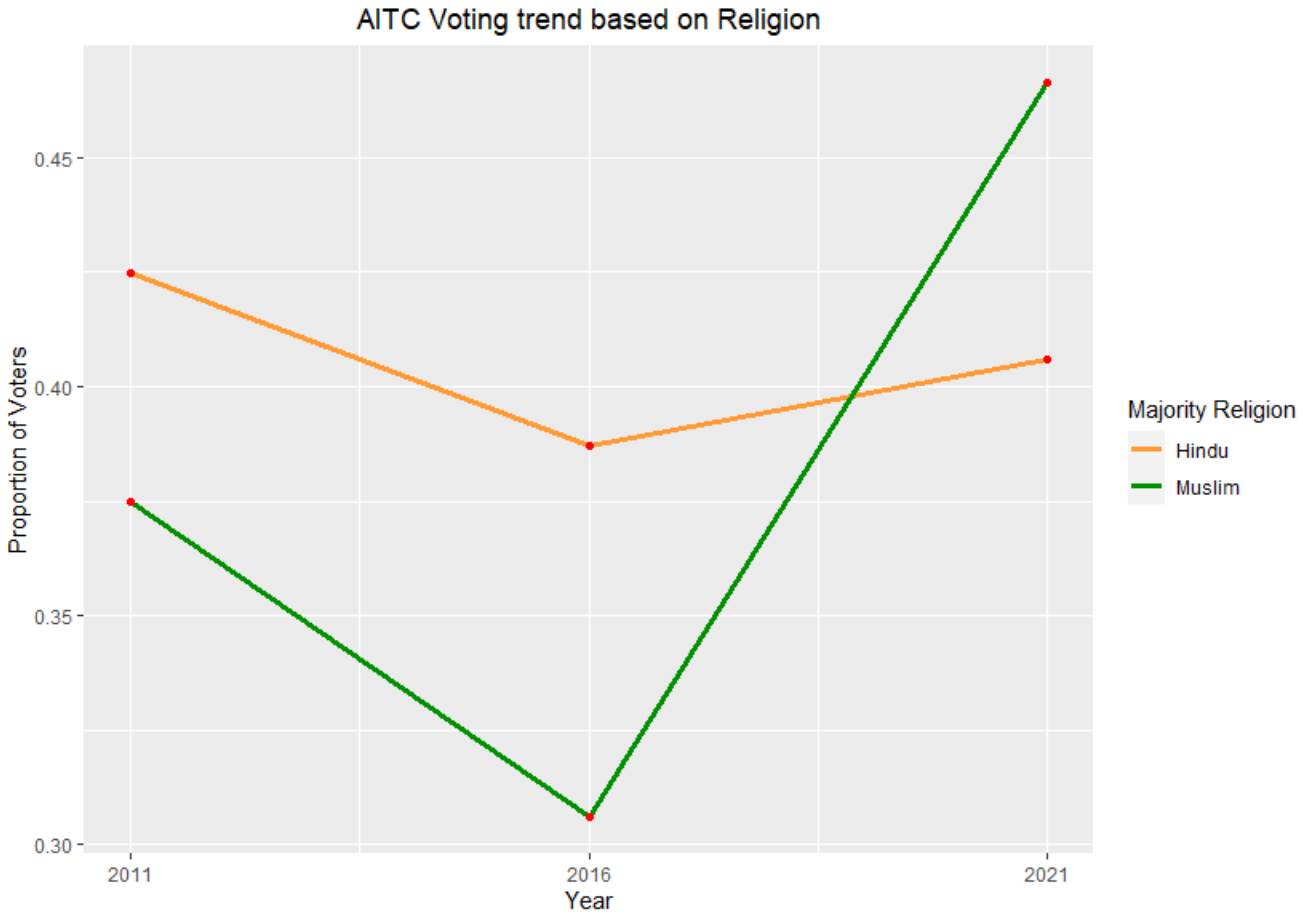
5 Trend Analysis

We see the trend of AITC votes over the years 2011, 2016 and 2021. We have chosen AITC because it won all three elections. Further, the votes of NDA and LF were inconsistent in the previous three elections. We first analyse based on majority religion and then based on the district. We see the average trend for these two categories in the last 3 elections.

5.1 Partition based on Religion

We partition the constituencies in two classes

- Hindu Majority areas: represented by the orange line.
- Muslim Majority areas: represented by the green line.



An important conclusion from here is the average vote share of AITC in Hindu dominant constituencies does not change drastically over the years, but that in Muslim dominant constituencies (52 out of 294) it has increased a lot.

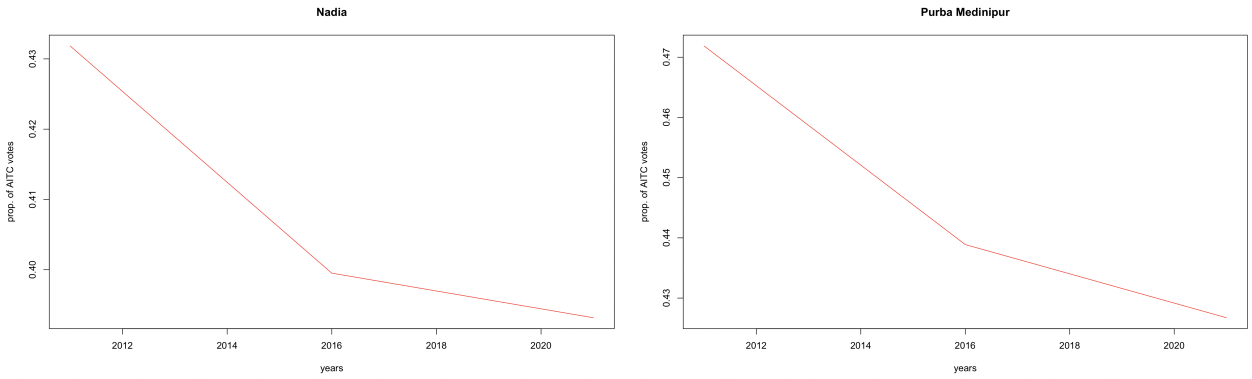
5.2 Partition on basis of Districts

Now we study the trend in the 23 districts of West Bengal. We classify them into 4 main categories on basis of the trend of proportion of AITC votes:

- A: Vote ratio strictly decreasing with year.
- B: Vote ratio peaked in 2016 but then fell down in 2021.
- C: Vote ratio attained minimum in 2016 and increased in 2021.
- D: vote ratio strictly increasing with year.

5.2.1 Category A

The districts in these category are Hooghly, Nadia, Paschim and Purba Medinipur. In these districts AITC vote share is on a steady decline.

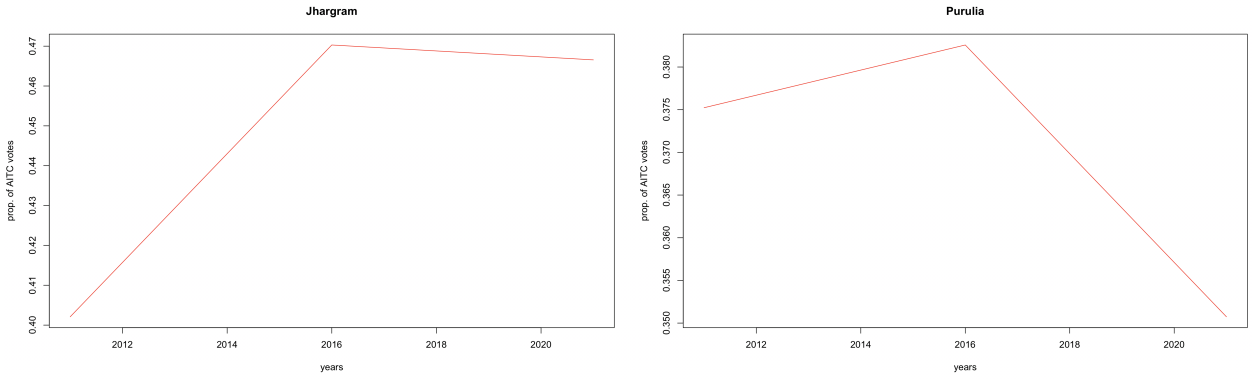


(a) Nadia (b) Purba Medinipur

Figure 1: Category A

5.2.2 Category B

The districts of Jhargram and Purulia fall under this category. Here AITC performed best in 2016. Although in Jhargram the decline in 2021 is minor, in Purulia it is more prominent.

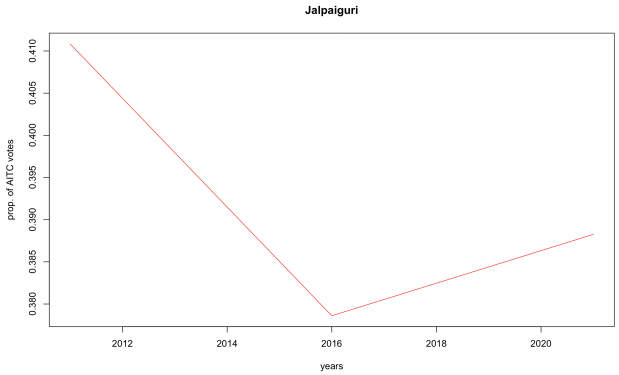


(a) Jhargram (b) Purulia

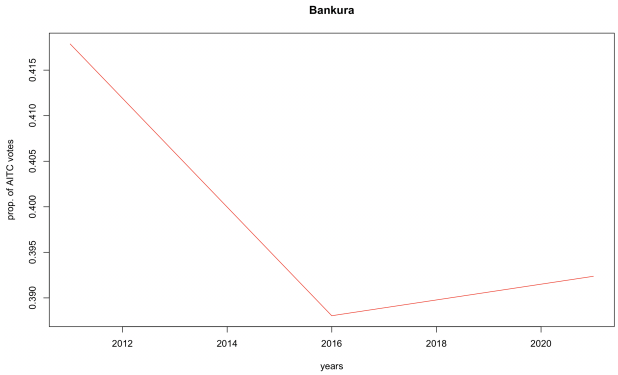
Figure 2: Category B

5.2.3 Category C

Most of the districts fall under this category. Here AITC had a dip in vote percentage in 2016 but did better in 2021. We further divide C into two categories, based on whether the maximum voter share was attained in 2011 or 2016. We call them low increase and high increase respectively. Among them Birbhum, Maldah, Murshidabad, Purba Bardhaman, South 24 Parganas and Uttar Dinajpur, AITC did much better than their last campaign (high increase). In Alipurduars, Bankura, Dakshin Dinajpur, Howrah, Jalpaiguri, Kolkata, North 24 Parganas and Paschim Bardhaman, the maximum was attained in 2011 (low increase).

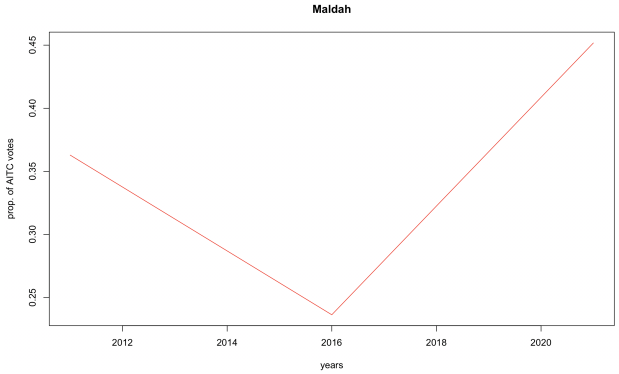


(a) Jalpaiguri

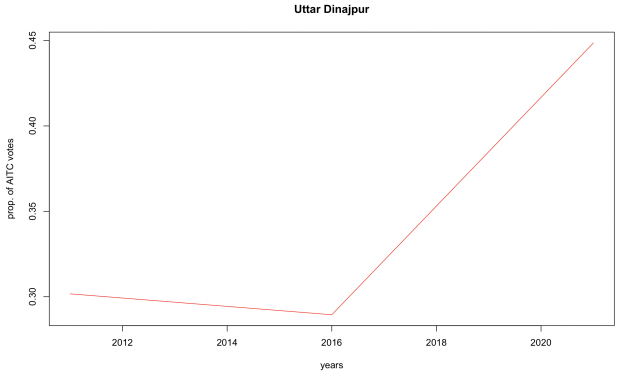


(b) Bankura

Figure 3: Category C(small increase)



(a) Maldah

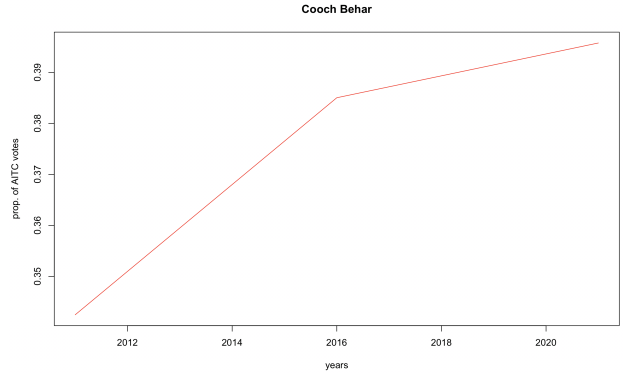


(b) Uttar dinajpur

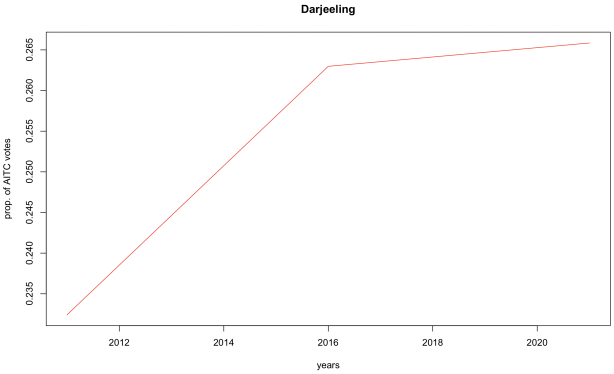
Figure 4: Category C(high increase)

5.2.4 Category D

Here AITC vote share have gone up with time. The districts are Cooch Behar and Darjeeling.



(a) Cooch Behar



(b) Darjeeling

Figure 5: Category D

The following table counts the number of constituency within each district of each category.

	Districts	A	B	C	D	Overall Category
1	Cooch Behar	2	2	2	3	D
2	Alipurduar	1	0	4	0	C
3	Jalpaiguri	0	2	4	1	C
4	Darjeeling	2	0	1	2	D
5	Uttar Dinajpur	0	0	6	3	C
6	Dakshin Dinajpur	1	0	5	0	C
7	Maldah	0	0	11	1	C
8	Murshidabad	0	0	17	3	C
9	Nadia	8	1	7	1	A
10	North 24 Parganas	9	3	16	5	C
11	South 24 Parganas	2	2	22	6	C
12	Kolkata Corporation	0	0	5	0	C
13	Kolkata	1	0	4	0	C
14	Howrah	1	3	11	1	C
15	Hooghly	8	2	8	0	A
16	Purba Medinipur	5	3	8	0	A
17	Paschim Medinipur	3	7	3	2	A
18	Jhargram	1	2	0	1	B
19	Purulia	2	2	3	2	B
20	Bankura	4	2	5	1	C
21	Purba Bardhaman	2	1	10	3	C
22	Paschim Bardhaman	0	0	8	1	C
23	Birbhum	0	3	6	2	C
	Total	52	35	166	38	

Most of the districts (and constituencies) were of type C. Purba and Paschim Medinipur are a bit unusual in the sense that even though they have most constituencies of type C and B respectively, the overall category is A. This can be explained by the fact that the C constituencies in these districts a show a small increase while A constituencies show a huge drop in AITC share. Hence, the overall category is A.

Here is a map of West Bengal with the districts coloured accordingly to the categories they belong to. It is observed that in the northern parts of West Bengal are category C or D hence AITC is essentially building a following here. However, in the South-Western and Nadia, parts their vote share is falling down a bit. We can further see that most districts are actually C.

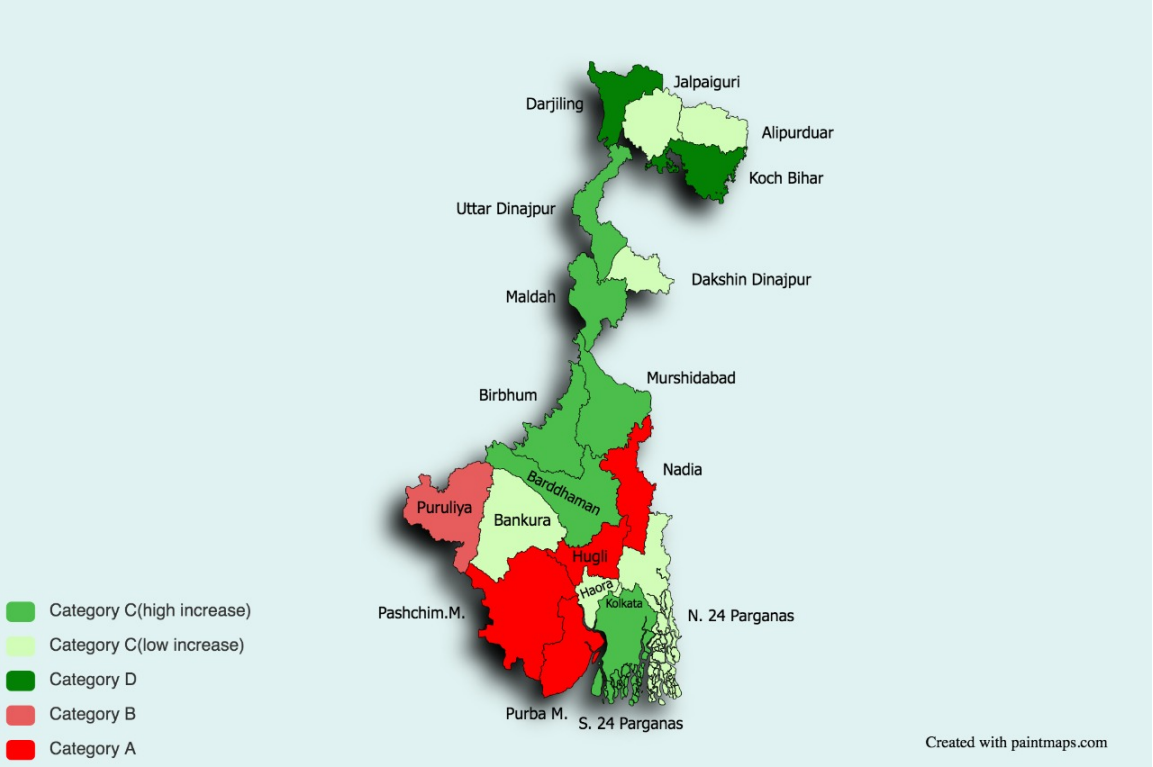


Figure 6: West Bengal

6 Regression Analysis

6.1 Simple Linear Regression

Notes Let D be a demographic class and A be a political party. Then (for a particular constituency)

$$\begin{aligned} \text{Votes}(A) &= \text{Votes}(A|D) + \text{Votes}(A|D^c) \\ \implies \mathbb{E}\left[\text{Votes}(A)\right] &= \mathbb{E}\left[\text{Votes}(A|D)\right] + \mathbb{E}\left[\text{Votes}(A|D^c)\right] \end{aligned}$$

Here, Votes(A) denotes the number of votes to party A in a particular constituency. Votes$(A|D)$ denotes the votes received by party A from people in demography D in a constituency. Expectation is over the population of a particular constituency since each individual behaves like a Bernoulli random variable who will vote in favour of A with some probability. Now, assuming that people with in a class behave alike (i.e. Block Voting assumption), we obtain that Votes$(A|D)$ and Votes$(A|D^c)$ must follows a Binomial Distribution. Further note that we actually know the values of Votes(A). Hence we can say

$$|D \cup D^c| \frac{\mathbb{E}\left[\text{Votes}(A)\right]}{|D \cup D^c|} = |D|P(\text{Vote for } A|D) + |D^c|P(\text{Vote for } A|D^c)$$

Replacing the expectation in LHS with the actual observed value, we get

$$\text{Proportion of Votes}(A) \approx \frac{|D|}{|D| + |D^c|} P(A|D) + \frac{|D^c|}{|D| + |D^c|} P(A|D^c)$$

or simply as a linear model

$$P(A) = \lambda P(A|D) + (1 - \lambda) P(A|D^c) + \epsilon$$

where λ is the proportion of a demography D. Note that we did not assume that everyone votes like a block. We only assumed people within a non trivial class vote as a block. Further, we know that value of $P(A), \lambda$ for all the constituencies. Hence, we can get the value of $P(A|D)$ and $P(A|D^c)$ using the method of Least Squares. Significance was checked using F test (details in Appendix). The normality of errors was checked using histogram. Further, we have a large sample so we don't need to worry about normality assumption too much. The exact implementation and p values has also been detailed in the Appendix.

Results Interesting results from 2011 (see table 1 on page 12). The model says that Hindus were more likely to vote for AITC than non hindus, Literate people were more likely to vote for AITC, Rural people were more likely to Vote for LF and voter base of other parties (other than AITC and LF) parties was mostly non Hindu and Illiterate. One unexpected result was that we obtained $P(\text{Other}|\text{Lit}) = -0.01$. While this is obviously not true, it suggests that most literate voters understood the concept of Tactical voting[1] since other parties did not have any chance in 2011. Another point to note is that Rural people were more likely to vote as compared to urban people. That is, Rural turnout was more.

2016 results (see table 2 on page 13) suggest that non Hindus moved from AITC to LF while a minor group Hindus moved towards other parties (most likely NDA) as compared to 2011. This year, the literate population in voter base of AITC and LF decreased and while the illiterate people voted more for LF. Again, Rural people were more likely to vote for LF and the turnout was low among urban population.

2021 (see table 3 on page 13) shows the most significant result of simple regression analysis. The NDA voter base was mostly hindus and 32.5% of the variation in NDA votes was explained by the Hindu population proportion. Further, this year illiterate people started voting for AITC while literate people were divided among AITC and NDA. Again, Urban population had a lower turnout ratio as compared to Rural. Another significant thing was that the literate population's turnout was much lower than illiterate's.

6.2 Restricted Regression

One problem with the previous model is that the estimated values of the probabilities are such that estimated $P(\text{AITC}|D)+P(\text{LF}|D)+P(\text{NDA}|D)+P(\text{Other}|D)+P(\text{NoTurnUp}|D)$ may not be equal to 1 for any demographic group D. This is clearly not a favorable situation, so we modified the least square method to incorporate the extra condition that the above-mentioned sum is 1.

Mathematically, we can model the situation as :

$$\begin{aligned}
\text{AITC}_i &= a_1 \cdot p_D^i + b_1 \cdot (1 - p_D^i) + \epsilon_i^{\text{AITC}} \\
\text{LF}_i &= a_2 \cdot p_D^i + b_2 \cdot (1 - p_D^i) + \epsilon_i^{\text{LF}} \\
\text{NDA}_i &= a_3 \cdot p_D^i + b_3 \cdot (1 - p_D^i) + \epsilon_i^{\text{BJP}} \\
\text{Other}_i &= a_4 \cdot p_D^i + b_4 \cdot (1 - p_D^i) + \epsilon_i^{\text{Other}} \\
\text{NoTurnUp}_i &= a_5 \cdot p_D^i + b_5 \cdot (1 - p_D^i) + \epsilon_i^{\text{NoTurnUp}}
\end{aligned}$$

subject to $\sum_{i=1}^{5} a_i = \sum_{i=1}^{5} b_i = 1$.
for the $i-$th constituency $(1 \leq i \leq 294)$ where X_i denotes proportion of votes won by X in constituency i and p_D^i denotes an estimate of the proportion of people of that constituency belonging to demography D. NoTurnUp refers to the proportion of electorate which did not go out to vote, and it can be envisioned as a separate party in this setup. Here, ϵ_i^X denote the error terms.

For compactness of expression , let $\text{AITC}_i = y_{1i}$, $\text{LF}_i = y_{2i}$, $\text{NDA}_i = y_{3i}$, $\text{Other}_i = y_{4i}$, $\text{NoTurnUp}_i = y_{5i}$. In order to obtain estimates for a_i's and b_i's, we minimize the SSE

$$\min_{a_j, b_j} \sum_{i=1}^{\#\text{cons.}} \sum_{j=1}^{5} (y_{ji} - a_j \cdot p_D^i - b_j \cdot (1 - p_D^i))^2$$

wrt the constraint. One downside of this model is that R^2 is no longer interpretable and F test no longer makes sense. The exact method is detailed in Appendix. The coefficients obtained are actually very similar to the coefficients obtained from simple regression. The largest difference was of 0.0265. The full tables can be seen on page 12.

[1] Tactical Voting means that people did not want to vote third party because they knew the third party had extremely low chance of winning

7 Comparing districts using Mahalanobis distace

The Mahalanobis distance is a measure of the distance between a point P and a distribution D. It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. This distance is zero for P at the mean of D and grows as P moves away from the mean along each principal component axis.

So let us say there are more than 2 parties and among them our interest is on the vote share of two parties A and B in all the districts. So we have a net vote share of A and B in West Bengal and we want to see if all the districts of West Bengal also follow the trend or some state of Bengal have conflicting interests with that of the state. (Note say 45% people in West Bengal prefer A and 30% prefer B then by outlier we not only look at districts where A performed relatively poorly but also districts where A won by a huge margin say got 80% of votes.)

The Mahalanobis distance of an observation $\vec{x}$ from a set of observations with mean $\vec{\mu}$ and covariance matrix S is defined as $D_i(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^{\mathsf{T}}\mathbf{S}^{-1}(\vec{x}-\vec{\mu})}$. Now we define $x_i = (x_1, x_2)_i$ where x_1 is the proportion of vote share of party A in district i and x_2 is the proportion of vote share of party B in district i. Note that the proportions are wrt the number of electors rather than the number of votes casted.

For 2011 and 2016, our observation is (Proportion of Votes of AITC, Proportion of Votes of LF). For 2021, the second observation was replaced by Proportion of Votes of NDA.

District	2011	2016	2021
Alipurduar	0.584	1.5029	1.5147
Bankura	0.812	0.7558	0.894
Birbhum	0.5075	0.7044	1.1196
Cooch Behar	1.8833	0.3921	1.3785
Dakshin Dinajpur	0.7781	0.8076	0.6088
Darjeeling	3.2569	1.1311	2.7608
Hooghly	1.2786	0.7297	0.0651
Howrah	0.9545	0.5747	0.4928
Jalpaiguri	0.3107	0.2094	1.0452
Jhargram	0.2633	2.8591	1.3418
Kolkata	2.5534	2.553	2.8277
Kolkata Corporation	2.2473	1.7816	2.0767
Maldah	0.9038	2.2552	1.1915
Murshidabad	0.617	2.2181	1.9428
Nadia	0.598	0.9542	0.6828
North 24 Parganas	1.2054	0.4958	0.6105
Paschim Bardhaman	0.2471	1.0888	1.2098
Paschim Medinipur	1.2712	1.2477	0.9743
Purba Bardhaman	1.3037	1.069	0.5004
Purba Medinipur	1.3634	1.5268	1.4049
Purulia	0.5174	0.5622	1.1047
South 24 Parganas	0.7964	0.7858	1.2774
Uttar Dinajpur	2.1343	1.2265	0.8886

The definition of outlier is distance > 2, i.e. point is at least two standard deviations away from the mean. The outliers have been marked in red for visibility . We will now explain **why** these are outliers :

- **Darjeeling 2011** was an extreme outlier , mainly due to the fact that 3 constituencies (out of 5) there elected GJM (an ally of BJP in 2011).
- **Kolkata/Kolkata Corp. 2011** Here TMC got about 1.91 times the votes of LF whereas in state they only got 1.19 times the votes of LF

- **Uttar Dinajpur 2011** was an outlier.It is because here LF won the election getting more than 5% extra votes than AITC.
- **Jhargram 2016** AITC got more than 63% of casted votes here which was quite high compared to the state.
- **Kolkata/Kolkata Corp. 2016** TMC in state had about 44% of casted votes in their favour but in Kolkata got more than 50%,along with that LF in state got more than 40% votes in state whereas in Kolkata got about 31% of casted votes.NDA vote share ratio was also much higher in Kolkata than what they got in the state.These makes the district an outlier when compared to the state.
- **Maldah 2016** was an outlier because LF obtained a huge number of votes and won 11 out of 12 constituencies there.
- **Murshidabad 2016** was an outlier, again, because LF dominated in this district.
- **Darjeeling 2021** NDA won here and NDA:AITC vote ratio was 1.46 which is too high compared to the state.
- **Kolkata/Kolkata Corp. 2021** AITC:NDA vote ratio in state was 1.26. But in Kolkata, AITC got more than twice the votes of NDA, so that is an outlier.

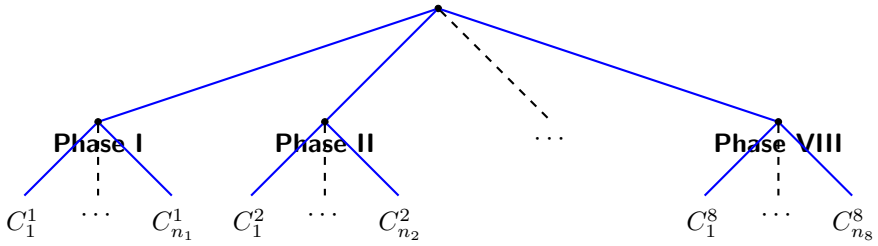
8 Do Demographies actually behave like a block?

We further wanted to check if demographies actually behave like a block or not. For this, we check if the interaction between demography and phase is relevant or not. First, we create an ANOVA model as follows

$$Y_i = \mu + \tau_{j(i)} + \epsilon_i$$

Here, i is the constituency number, Y_i is the proportion of votes a party got and $j(i)$ is the phase number during which the elections of constituency i were held.

The hierarchy of this setup is as follows :



Note that the election in 2021 was held in 8 phases . In the diagram, $C_1^j, \cdots, C_{n_j}^j$ represent the n_j-many constituencies where elections were conducted in the j-th phase.

We further tested if the phase effect was significant or not. It turned out phase effect was significant for both NDA and AITC with p values of 10^{-12} and 0.0085 respectively. We hypothesized that there is a "base" probability for each phase and then there is an additional probability if the person is from a particular demographic. Essentially, we compared the following two models.

$$Y_i = \mu + \tau_{j(i)} + \beta_{j(i)} D_i + \epsilon_i$$

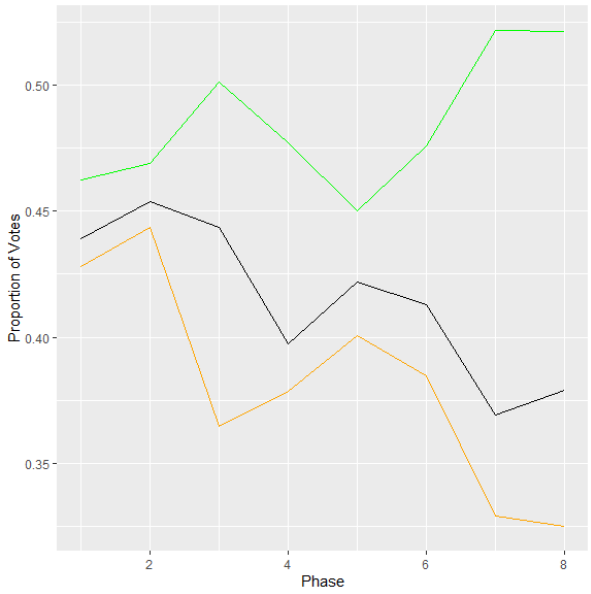
$$Y_i = \mu + \tau_{j(i)} + \beta D_i + \epsilon_i$$

Since the first model was a good fit, we did a likelihood ratio test too see if those parameters are actually needed or not. For NDA and Hindus, we get a p value of 0.5 which suggest that the extra parameters are not needed. The values of β was 0.22 which means that compared to non Hindus, Hindus voted for NDA with 22% more probability and this probability was consistent throughout the phases. Analysis also suggested that this is not true for literate people. Literate people behaved differently in different phases. But this was true for Urban population, with urban population having a probability modifier of -4% for voting for NDA.

As for AITC, block voting assumption may not be a very good choice. The null hypothesis was rejected (with all three explanatory demographies) with the highest p value being 0.017. A possible explanation for this is that AITC has been the incumbent party for 10 years, hence people in different location have different opinion depending on the work done by AITC but since BJP has never been in power WB, people throughout Bengal have more or less a similar opinion of BJP.

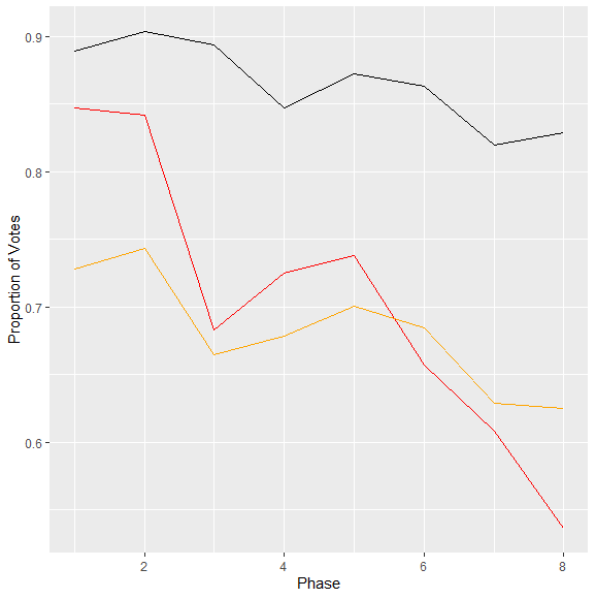
9 Turnout Dependence

Next we tried to see if turn out proportion is related to proportion of votes received by NDA and AITC. Initially, we tried to fit a linear model. Though the QQ plot was fairly good, the errors were not heteroscedastic. Hence, we partitioned the constituencies based on the Phase of the elections to check graphically is there is any association. On the right attached is a graph which shows proportion of phase wise turnout (black), proportion of NDA vote (Orange) and AITC votes (green). Note that the y axis is **NOT** correct for the turnout! All points of the black line have been reduced by 0.45 to enable easy comparison. So turnout in Phase 2 was actually 0.904 and turnout in Phase 4 was actually 0.85.



On the first glance, it looks like lower turnout means lower proportion of votes for NDA and vice versa for AITC votes. However, it still remains to be checked if there is any confounding effect or not. This elections, the phases were such that proportion of Hindus decreased with phase. In the latter phases, the turnout, the proportion of Hindus and the proportion of NDA votes, all three decrease simultaneously. This can be seen in the graph below. We also calculated the Pearson correlation coefficient for proportions of Hindus and turnout ratio, it turned out to be -0.05 which is almost insignificant.

Most of Maldah, Uttar Dinajpur and Murshidabad (Muslim majority districts) had their election in last 3 phases making proportion of Hindu voters lower as election gradually proceeded. Now as phases proceeded from March 27 to April 29 India started to be more and more affected by Covid each passing day. It's possible that Covid fears had forced people to stay indoors, so with the passing of time across subsequent phases, the turnout in urban part of Bengal reduced considerably. But it was not the same in rural Bengal as in most districts it was around 80%-plus voting. Lower turnout generally favours incumbent party as we saw and along with that the decline in the proportion of Hindus are some of the probable causes for fall of NDA vote share in the later stages of the election.



10 Appendix

Manual Change in data for 2016 election Nihar Ranjan Ghosh of English Baazar was an independent candidate. His party was changed to CPI in the data since he allied with them.

Implementation of Simple Linear Regression
The model given in Section 6.1 is

$$Y_i = \alpha X_i + \gamma(1 - X_i) + \epsilon_i$$

where i denotes the constituency number, Y denotes the proportion of votes to a particular party and X denotes the proportion of people from a particular demography. Notice that there is no mean μ term hence, a single row of model matrix looks like $[Hn,\ (1 - Hn)]$ rather than $[1,\ Hn,\ (1 - Hn)]$. Hence, the regressors are not collinear. However, actually the following model has been implemented in R to make analysis easier.

$$Y_i = \mu + \beta X_i + \epsilon$$

This is because the standard F test in R relies on μ being a parameter. It is easy to see that the column space of the regressor remains the same. The relation between the coefficient of the two models is given by $\gamma = \mu$ and $\alpha = \mu + \beta$. This can be seen by grouping the coefficient of X_i together in first model. The null hypothesis for F test in the second model is $\beta = 0$ which translates to $\alpha = \gamma$ in the initial model. Hence, the p-values in the tables below signify if the difference between the two estimates (α and γ) is non zero or not.

Outliers in Regression analysis

- 2011: Kalimpong, Darjeeling and Kurseong were extreme outliers hence removed from analysis (dominated by GOJAM which allied with BJP) . Also, Joypur, Malatipur, Dinhata, Chopra were outliers with less than 15% AITC votes.
- 2016: Rejinagar was an outlier this year because of less than 8% AITC votes.
- 2021: Bhangore, Chowrangee, Jorasanko were outliers for 2021. Bhangore was an outlier because of high percentage (47.77 %) votes to other independent candidates/parties. Chowrangee and Jorasanko were outliers because of low turnout as compared to the state ($< 60\%$)

Simple Regression Tables

Below are the tables for simple linear regression. P value is derived from F test which tests if $P(A|D) = P(A|D^c)$. P Value less than 5% denotes that $P(A|D) \neq P(A|D^c)$

Table 1: Simple Regression in 2011

	AITC	LF	Other	NoTurnUp
Hn	0.4542	0.3544	0.042	0.1494
Non Hn	0.353	0.3476	0.1431	0.1563
P Value	0	0.7209	0	0.7143
R sq	0.1189	0.0004	0.134	0.0005
Lit	0.5188	0.3156	-0.0091	0.1747
Illit	0.2314	0.4269	0.2372	0.1044
P Value	0	0.0055	0	0.0742
R sq	0.2157	0.0267	0.1786	0.0111
Ub	0.4372	0.3036	0.0544	0.2048
Rural	0.4179	0.3741	0.0803	0.1278
P Value	0.0299	0	0.0019	0
R sq	0.0164	0.1796	0.0332	0.2249

Table 2: Simple Regression in 2016

	AITC	LF	Other	NoTurnUp
Hn	0.4151	0.2949	0.1235	0.1673
Non Hn	0.2772	0.4153	0.1309	0.1793
P Value	0	0	0.6995	0.5206
R sq	0.1152	0.0946	0.0005	0.0014
Lit	0.4578	0.2538	0.0781	0.2125
Illit	0.2038	0.4861	0.2211	0.0869
P Value	0	0	0.0003	0.0011
R sq	0.0897	0.08	0.0457	0.0358
Ub	0.3517	0.287	0.1411	0.2234
Rural	0.3832	0.3506	0.119	0.1481
P Value	0.0099	0	0.0241	0
R sq	0.0227	0.0984	0.0179	0.2122

Table 3: Simple Regression in 2021

	AITC	NDA	Other	NoTurnUp
Hn	0.3771	0.4126	0.0691	0.1406
Non Hn	0.5115	0.1397	0.2222	0.127
P Value	0	0	0	0.5056
R sq	0.1456	0.3253	0.2274	0.0015
Lit	0.3619	0.3736	0.0795	0.1864
Illit	0.5289	0.245	0.1865	0.0358
P Value	0.0001	0.0287	0.0065	0.0003
R sq	0.0519	0.0166	0.0256	0.0441
Ub	0.3876	0.3035	0.1149	0.194
Rural	0.4297	0.3426	0.1149	0.1124
P Value	0.0001	0.0069	0.9958	0
R sq	0.0539	0.0252	0	0.2114

Restricted Multiple Regression

Below are the results for restricted regression.

Restricted Regression Model Implementation Will given an example with three parties, A, B and C. The non restricted model for a particular constituency is given by

$$\begin{bmatrix} A_i \\ B_i \\ C_i \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & D_i & 0 & 0 \\ 0 & 1 & 0 & 0 & D_i & 0 \\ 0 & 0 & 1 & 0 & 0 & D_i \end{bmatrix} \begin{bmatrix} \mu_A \\ \mu_B \\ \mu_C \\ \beta_A \\ \beta_B \\ \beta_C \end{bmatrix} + \vec{\epsilon}$$

It is fairly easy to see how to extend it to multiple constituencies. This can be written compactly as

$$\begin{bmatrix} A_i \\ B_i \\ C_i \end{bmatrix} = \left(\begin{bmatrix} 1 & D_i \end{bmatrix} \otimes \mathbb{I}_3 \right) \beta + \vec{\epsilon}$$

We wish to restrict the model so that

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \beta = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

which is same as

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \beta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The first restriction says that $P(A|D^c) + P(B|D^c) + P(C|D^c) = 1$ and the second restriction says that $P(A|D) + P(B|D) + P(C|D) = 1$. This restriction can be compactly written as

$$\left(\mathbb{I}_2 \otimes \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}\right) \beta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Since we're making two restrictions on a model with 6 parameters, our new model will have 4 parameters. Call the vector of these 4 parameters to be α. Further, α must be such that β can be uniquely determined given a values of α and the restrictions. Hence, $\alpha = \begin{bmatrix} \mu_A & \mu_B & \beta_A & \beta_B \end{bmatrix}^T$ is a valid choice while $\alpha = \begin{bmatrix} \mu_A & \mu_B & \mu_C & \beta_A \end{bmatrix}^T$ is not. For the former value of α, it is easy to see that

$$\beta = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & -1 \end{bmatrix} \alpha + \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

which can be written as

$$\beta = \left(\mathbb{I}_2 \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}\right) \alpha + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

plugging in this value of β in the original model, we get

$$\begin{bmatrix} A_i \\ B_i \\ C_i \end{bmatrix} = \left(\begin{bmatrix} 1 & D_i \end{bmatrix} \otimes \mathbb{I}_3\right) \left(\left(\mathbb{I}_2 \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}\right) \alpha + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \otimes \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) + \vec{\epsilon}$$

Simplifying using the properties of outer product, we get

$$\begin{bmatrix} A_i \\ B_i \\ C_i \end{bmatrix} = \left(\begin{bmatrix} 1 & D_i \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}\right) \alpha + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} + \vec{\epsilon}$$

This gives use the model that we want to fit by

$$\begin{bmatrix} A_i \\ B_i \\ C_i - 1 \end{bmatrix} = \left(\begin{bmatrix} 1 & D_i \end{bmatrix} \otimes \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{bmatrix}\right) \alpha + \vec{\epsilon}$$

However, this model has a few downsides. For one, the usual interpretation of R^2 is no longer valid.

Table 4: Restricted Regression in 2011

	AITC11	LF11	Other11	NoTurnUp11
Hindus	0.4542	0.3544	0.0419	0.1495
Non Hindus	0.353	0.3476	0.1431	0.1563
Lit	0.5202	0.3046	-0.002	0.1772
Illlit	0.2049	0.4391	0.2521	0.1039
Ub	0.4372	0.3037	0.0544	0.2047
Rural	0.4179	0.3741	0.0803	0.1277

Table 5: Restricted Regression in 2016

	AITC16	LF16	Other16	NoTurnUp16
Hindus	0.4149	0.2947	0.1233	0.1671
Non Hindus	0.2765	0.4146	0.1302	0.1787
Lit	0.4572	0.2533	0.0776	0.2119
Illlit	0.2043	0.4866	0.2217	0.0874
Ub	0.3509	0.2862	0.1403	0.2226
Rural	0.383	0.3504	0.1188	0.1478

Table 6: Restricted Regression in 2021

	AITC21	NDA21	Other21	NoTurnUp21
Hindus	0.3773	0.4128	0.0693	0.1406
Non Hindus	0.5114	0.1397	0.2221	0.1268
Lit	0.3616	0.3732	0.0791	0.1861
Illlit	0.5298	0.2459	0.1874	0.0369
Ub	0.3876	0.3035	0.1149	0.194
Rural	0.4298	0.3427	0.115	0.1125

References

[1] West Bengal Chief Electoral Office. *Constituency wise elector data*. URL: `http://ceowestbengal.nic.in/UploadFiles/PE2019/PC_Dist_AC_PS_PSL.pdf`. (accessed: 05-06-2021).

[2] Registrar General and Census Commissioner of India. *Census 2011*. URL: `https://censusindia.gov.in/2011census/population_enumeration.html`. (accessed: 01-04-2021).

[3] Election Commision of India. *WBLA Results 2011*. URL: `https://eci.gov.in/statistical-report/statistical-reports/`. (accessed: 01-04-2021).

[4] Election Commision of India. *WBLA Results 2016*. URL: `https://eci.gov.in/statistical-report/statistical-reports/`. (accessed: 01-04-2021).

[5] Election Commision of India. *WBLA Results 2021*. URL: `https://results.eci.gov.in/Result2021`. (accessed: 04-06-2021).