**Write-up on the project – Prabhat Singh**

**Executive Summary**

My project is based on data from the Indian national elections concluded recently. There are a total of 542 constituencies in the country, and the project scrapes different types of data, from the website of Election Commission of India, for each of those constituencies.

A total of five variables have been scraped for each constituency: name, margin of victory as measured by number of votes, number of candidates that contested, total votes polled, and the party to which the winning candidate belongs.

The goal of the project is to determine the competitiveness of the elections, which is measured by the margin of victory in each constituency. Given the limited scope of the project, I have only included total votes polled and number of candidates that contested, as the independent variables.

**Details**

The 6 constituencies of the state of Jammu & Kashmir have been dropped because of the anomalous structure of the page, which made scraping difficult. This leaves a total of 536 constituencies for the analysis.

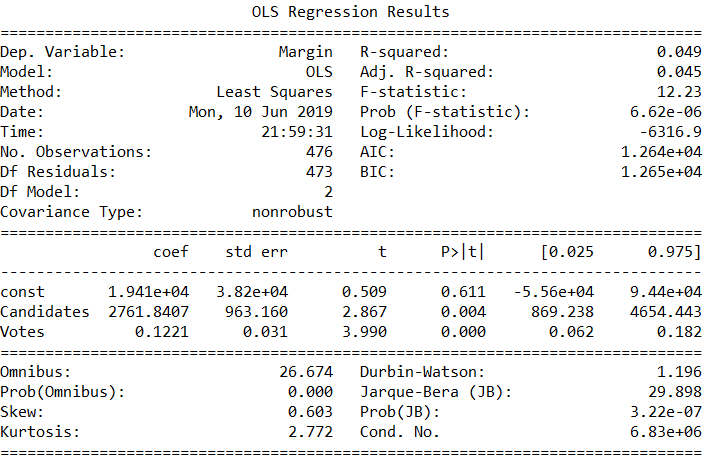
The choice of these two variables – total votes polled and number of contesting candidates – was obvious. Intuitively, the margin of victory should go up as more votes are polled, and go down as more candidates enter the fray, since votes are expected to be divided up between them.

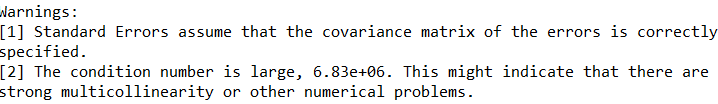
I tried figuring out the dependence of margin of victory on the party of the winning candidate by including it as a dummy variable, but chose to drop this from the final analysis given the limitations of the scope of the project as well as of my own statistical knowledge. Ideally, some sort of ordinal regression that ranks the parties – highest rank going to the party that has the largest impact on the margin of victory – should be performed.

**Results**

My model does not show robust results (picture below). The R-squared is very low, thereby predicting a poor fit. The regression summary makes it clear that there’s a high degree of multicollinearity. The constant is way too high, and so is the coefficient of the variable Number of Candidates. Both the variables are statistically significant, given the low p-values.

Furthermore, the model throws up warnings of multicollinearity.





Undoubtedly, there is omitted variable bias in my analysis. A more accurate and large scale analysis must include the party of the winning candidate. There are several other factors that my model doesn’t take into account.

**Further discussion**

Some of the papers that might be relevant to my model are:

1. <https://www.journals.uchicago.edu/doi/abs/10.2307/2131719>

2. <http://web.stanford.edu/~jgrimmer/CEF.pdf>

Ideally, this model must include, at the very least, the money spent by the winning candidate (and his/her party) in a given constituency. This would arguably be the most important variable predicting margin of victory. However, this data is unreliable and very hard to obtain, so I did not include it. Other factors could be whether the winning candidate is an incumbent or not, performance of the incumbent, a measure of the “popularity” of the candidate etc. Number of criminal cases against a candidate would also be an interesting variable to include, given the increasing proportion of winning candidates in India who have criminal cases against them. It is quite possible that margin of victory is positively correlated with number of criminal cases.

The complexity with including the above variables is that they’re not independent of each other. For instance, the money spent is highly likely to be correlated with the party of the candidate, as larger parties tend to have deeper pockets. It would be up to researchers to figure out how to deal with this issue.

**Conclusion**

Given the limitations of my model, it would be an interesting and worthwhile exercise to model margin of victory on a greater number of factors, some of which have been named above.

In addition, there is scope for comparison of the competitiveness of latest elections with those of previous elections (large amounts of data is available to facilitate this comparison), to check for change in competitiveness as well as for factors influencing competitiveness that were relevant then and now.