## **Introduction**

Voter turnout is crucial in American politics. Democracy and civic discourse are impotent without voter turnout, and checks and balances on the political system are enforced at the ballot box. Campaigns have strong incentives to mobilize voter turnout for their candidates. Predicting voter turnout, however, is challenging. The United States’ relatively low voter turnout compared to other democracies complicates the use of polls because pollsters and census takers may not be sampling the “right population,” i.e. -- those who will actually vote. One solution to this problem is to poll registered voters; however, not all registered voters can be counted on to actually vote. In order to solve this problem, pollsters can build likely voter models to estimate the population of those who will actually vote.

Pollsters who wish to use such models must overcome two challenges. The first challenge is limitations with the models themselves, and the second is limitations with human capital. Ordinary least squares (“OLS”) and logistic regression are effective tools in constructing likely voter models. When combined, they use variation in characteristics of voters to predict the likelihood of voting. However, OLS is quite sensitive to issues of collinearity among predictor variables, and characteristics associated with voting behavior are likely to be highly correlated. Little to no collinearity among predictors is a key assumption of logistic regression. An expert analyst can mitigate issues of collinear predictors through model selection, but these solutions are not often generalizable beyond a given locality and year and require significant domain knowledge to implement. Campaigns and pollsters faced with limited resources may be unable to attract analysts capable of building such models and interpreting their output.

To overcome these challenges, this paper builds a model using principal components logistic regression to predict voter turnout using data collected from people registered to vote in recent general and mid-term elections. We intend to answer the question of whether one can build a general (i.e. – national-level) model that predicts voter turnout and is robust to significant collinearity in the predictors with limited judgement required in its interpretation.

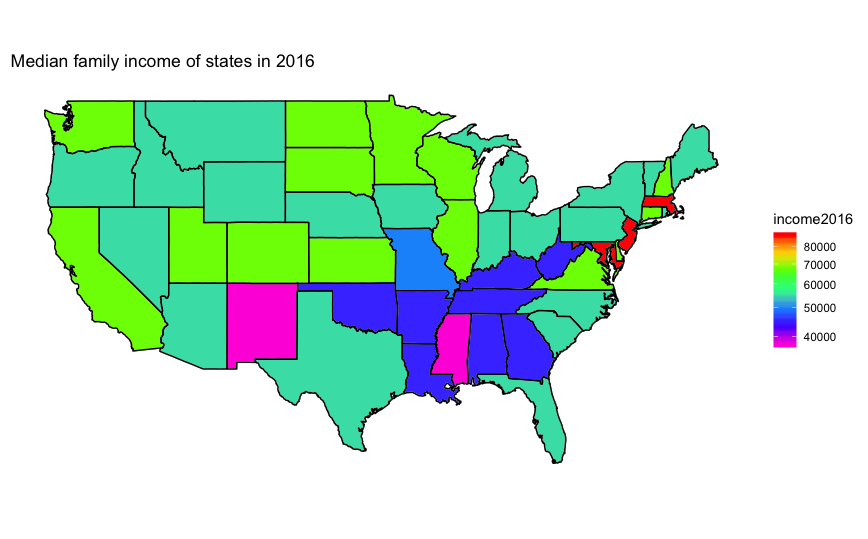
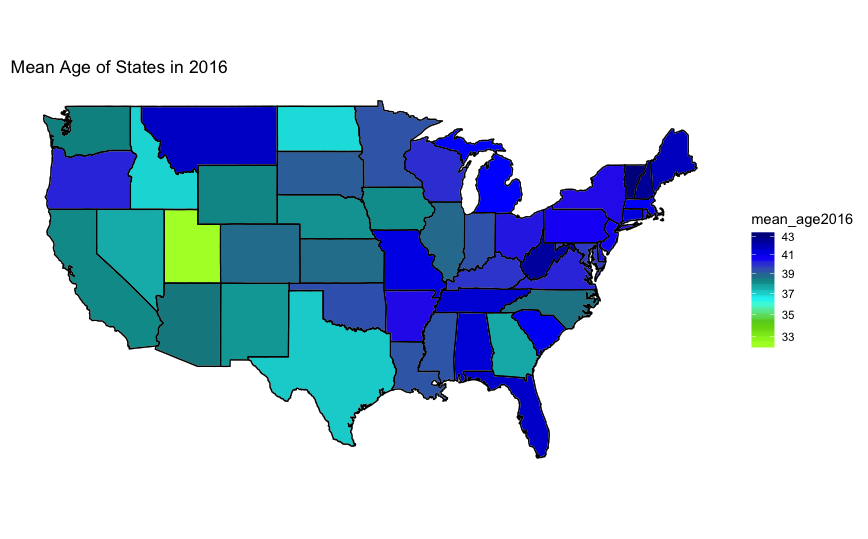
## **Data & Methods**

This section begins with a description of our data and moves to a discussion of our method of analysis. Our data are taken from the Voter Supplement of the Current Population Survey (“CPS”) administered jointly by the US Census Bureau and the US Bureau of Labor Statistics (“BLS”) immediately following every general and mid-term election and made available by the Integrated Public Use Microdata Series (“IPUMS”) at the University of Minnesota one month after each election. Because historical voting habits are a good indicator of future voting habits, this data offers a rich resource of demographic and prior voting habits for constructing a model that will enable us to predict whether the next registered voter is likely to vote.

Our sample is restricted to elections beginning in 1994 to the present. We begin here because this is an inflection point in the habits and preferences of American voters. During the Clinton presidency, certain groups of voters became galvanized over political and cultural touchstones, and Newt Gingrich, who became Speaker of the House in 1994, spearheaded a trend toward a new, more partisan, “warfare” politics. Therefore, voting behavior prior to 1994 may not be predictive of voting behavior post 1994. We further restrict our sample to registered voters who recorded a response of “Voted” or “Didn’t Vote.” This excludes individuals who did not respond or who were ineligible to respond.

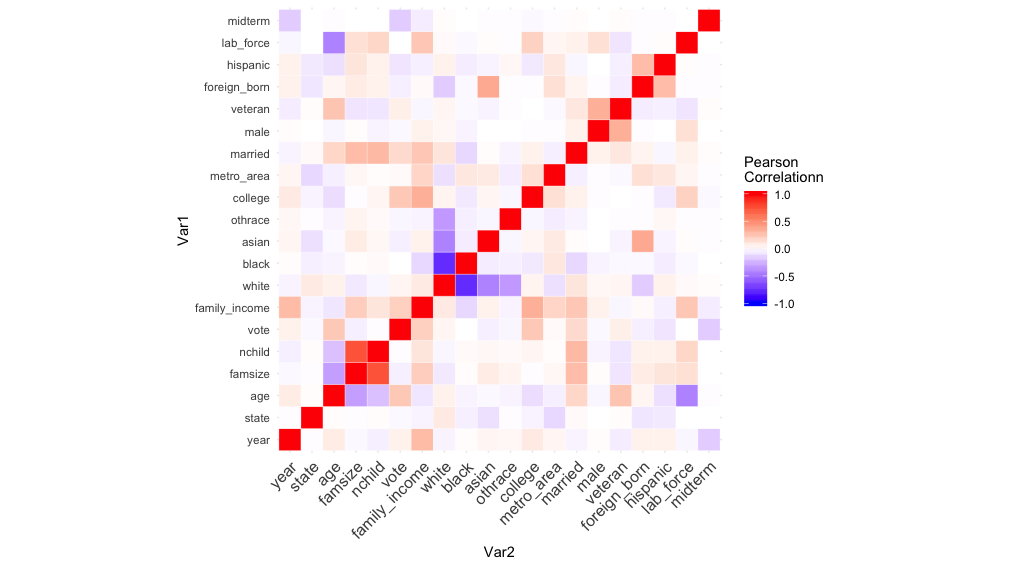
We present summary statistics for all variables and graphics for two continuous variables: mean age and median family income by state. We omit median values for binary variables because they offer no interpretation. The variable *college* is an indicator for whether someone in the sample has attended any college, whether the person has completed college or not. The variable *metro\_area* is an indicator for whether someone in the sample lives in a metropolitan area. There appears to be a large amount of variation in all variables in our sample.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Summary Statistics (N = 736,829)** | | | | | |
|  | Mean | SD | Median | Min | Max |
| age | 47.122 | 17.566 | 46 | 18 | 90 |
| famsize | 2.771 | 1.488 | 2 | 1 | 16 |
| nchild | 0.744 | 1.093 | 0 | 0 | 9 |
| vote | 0.620 | 0.485 |  | 0 | 1 |
| family\_income | 54,987 | 38,689 | 45,000 | 2,500 | 150,000 |
| white | 0.842 | 0.365 |  | 0 | 1 |
| black | 0.101 | 0.301 |  | 0 | 1 |
| asian | 0.034 | 0.182 |  | 0 | 1 |
| othrace | 0.023 | 0.149 |  | 0 | 1 |
| college | 0.566 | 0.496 |  | 0 | 1 |
| metro\_area | 0.709 | 0.454 |  | 0 | 1 |
| married | 0.597 | 0.490 |  | 0 | 1 |
| male | 0.471 | 0.499 |  | 0 | 1 |
| veteran | 0.117 | 0.322 |  | 0 | 1 |
| foreign\_born | 0.074 | 0.261 |  | 0 | 1 |
| hispanic | 0.072 | 0.258 |  | 0 | 1 |
| lab\_force | 0.676 | 0.468 |  | 0 | 1 |
| midterm | 0.51 | 0.50 |  | 0 | 1 |

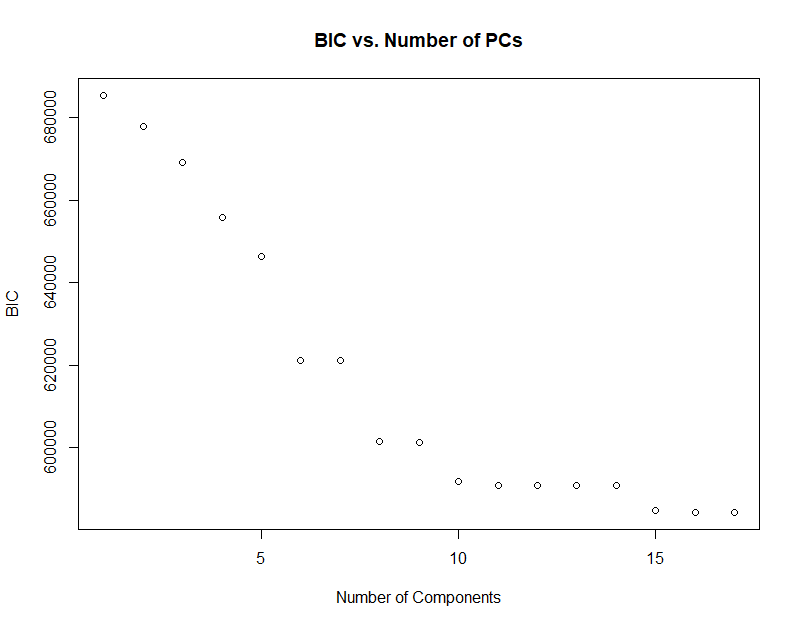


There is a high number of categorical variables in our data. Many of these are recoded to binary variables, such as categories for race. Family income is recoded from categorical to numeric by imputing the mean of each category. We omit all observations which have missing data as is necessary for principal component analysis. This reduces our sample from 1,596,223 to 736,829. Since our method dictates that we have complete information necessary to predict voting behavior, we accept that this may cause issues for the representativeness of our sample.

This paper uses principal components logistic regression (“PCLR”) to predict the likelihood of voting based on demographic predictors and historical voter turnout. We make use of logistic regression because we are predicting a binary outcome. We use principal components regression (“PCR”) because the predictors are likely highly correlated, and the dimensionality reduction properties of PCR will help isolate predictive variation among collinear predictors. The following correlation matrix supports this hypothesis. Many variables are indeed correlated with each other in the neighborhood of 0.5 or greater in magnitude.



Although PCR is more commonly employed in unsupervised contexts, the known outcome of whether someone votes makes this a supervised learning application. We therefore estimate PCLR as our primary model and compare our performance with a simple logistic regression model based on a 0-1 loss function with number of children dropped due to high collinearity with family size and the race indicator “white” as a reference category. We will assess model performance based on a mis-classification rate, which is the number times our model produces an incorrect result divided by the total number of test observations. We now detail how we use cross-validation and bootstrapping.

We begin with our use of cross-validation. We first randomly shuffle the data and then separate it equally into five folds for both the principal component transformed data frame and the standard data frame using the same seed. We then estimate the mean five-fold cross-validation error for the standard logistic regression and iterate for each successive number of principal components. We then compare the cross-validated mean mis-classification rate from the standard logistic model to the PCLR as each principal component is added. The goal is to identify the number of principal components necessary to achieve a mean mis-classification rate comparable to standard logistic regression.

We move now to bootstrap. The typical process for bootstrap error estimation can result in overlap between the training and test data if one uses the bootstrapped sample as either the training or test sample. This results in the model training on some of the very same data on which it will eventually be tested. To sidestep this issue, we first split both the transformed data and standard data into training and test segments, 70% and 30%, respectively. First, using only the training data, we fit a logistic model with each successive number of principal components and examine the BIC as each component is added to identify candidate models for comparison to standard logistic.

Upon examining the BIC, we see that large decreases occur at six, eight, and 10 principal components. We therefore compare PCLR models using six, eight, and 10 principal components with the standard logistic regression. We perform this screening process because it would be computationally infeasible to estimate models for each number of principal components. A smaller but noticeable change in BIC occurs between 14 and 15 principal components. However, we do not use these because they would approach the same estimates as the standard logistic, and this would eliminate the benefits of dimensionality reduction.

Having settled on six, eight, and 10 principal components as candidate models for comparing the PCLR to the standard logistic, we move on to the next step in applying the bootstrap. We generate each bootstrap sample using only the training data for both the principal component and standard data frames. In each iteration, we refit all models and produce the test error using the non-boostrapped test data. We iterate this process 100 times and average the mis-classification rate for each of the four models. We then compare the mis-classification rates of the three PCLR models with that of the standard logistic model. We make use of the “readstata13” packages for importing the data. The “stats” and “mass” libraries are used for various functions. We reproduce our code in full in an appendix.

# **Results**

This section presents and compares the results of our PCLR and logistic regression analyses. Table 1 presents our results for cross-validation.

|  |  |
| --- | --- |
| Table 1: Mis-classification Rates for Cross-validation | |
| Number of Principal Components | CV Misclassification Rate |
| 1 | 0.38032 |
| 2 | 0.38706 |
| 3 | 0.39122 |
| 4 | 0.36541 |
| 5 | 0.34749 |
| 6 | 0.32540 |
| 7 | 0.32554 |
| 8 | 0.30954 |
| 9 | 0.30943 |
| 10 | 0.30065 |
| 11 | 0.30030 |
| 12 | 0.30013 |
| 13 | 0.30009 |
| 14 | 0.29991 |
| 15 | 0.29469 |
| 16 | 0.29418 |
| 17 | 0.29419 |
| Standard Logistic | 0.29430 |

Our stated goal is to produce results that are at least as good as standard logistic regression. The standard logistic regression has a mis-classification rate of 0.29430, and comparable mis-classification from the PCLR model occurs at 15 principal components. The following graph illustrates the results in Table 1 with the standard logistic mis-classification in red indicated by the red line.

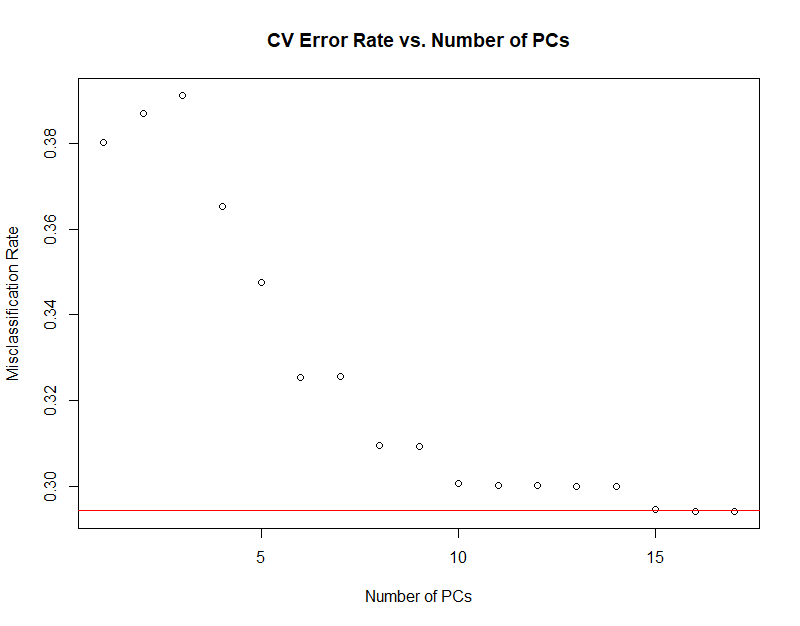


Table 2 lists the mis-classification rates for the standard logistic model and each of the three PCLR candidate models.

|  |  |
| --- | --- |
| Table 2: Mis-classification Rates for Bootstrap | |
|  | Misclassification Rate |
| Standard Logistic | 0.31693 |
| 6 PCs | 0.32416 |
| 8 PCs | 0.30892 |
| 10 PCs | 0.29975 |

The eight- and ten-component models perform better than the standard logistic based on our metric while the six-component model underperforms.

We move on to a discussion of model assumptions. There are only a few assumptions for logistic regression, namely: independent observations and little or no collinearity among predictors. Since we are using repeated cross-sectional data rather than panel data, we do not expect serial correlation to be a problem. However, errors may be correlated among observations that share similar characteristics, such as living in the same state or voting in the same year. Since we are not examining statistical significance of individual predictors, this is not a problem.

We reiterate that a benefit of using PCR is to neutralize the collinearity in the predictor space, and therefore, the assumption of collinearity is satisfied by the orthogonal predictors generated via principal components. A final, trivial assumption imposed by the logistic regression model is that our predictors are linearly associated with the log-odds of our outcome. This is merely an artifact of the model used to create a smooth distribution in the outcome, and we have no way of addressing the veracity of this assumption.

# **Discussion & Conclusion**

Based on our results, cross-validation suggests that we need 15 principal components to achieve a mis-classification rate similar to the standard logistic regression, while bootstrapping suggests that we only need eight. There is a clear trade off between prediction accuracy and the number of components between the two methods. By bootstrapping, we converge to the standard logistic mis-classification rate with fewer principal components.

To provide some insight into interpreting this, we revisit our initial step in the bootstrap: the train test-split. By randomly dividing the data and bootstrapping only the training segment, we make the bootstrap sensitive to the observations which happen to be in the training segment. This creates two problems. First, it is unclear that the bootstrap is gaining much over a simple train-test split without bootstrapping the training segment. Second, if there are outliers in the training data, bootstrapping the sample will amplify the effects of those outliers.

A possible solution is to iterate the process of segmenting the data 100 times, and we may see a different rate of convergence based on the observations in the individual training samples at each initial step in the iteration. However, this is computationally demanding, and it is unclear what a more favorable outcome would be in this context except to give us more confidence in the result. Furthermore, this approach seems quite close to repeated cross-validation, but more complicated and with no clear advantages.

Repeated cross-validation is another approach that may yield some insight. We could have repeated the five-fold cross-validation 100 times and taken an average over those 100 iterations. This process is less sensitive to the observations in each sample since it is performed without replacement. However, we again run into problems of computational feasibility.

In closing, based on the results above we are confident that our model would deliver comparable predictive ability to that of a logistic regression used by a trained analyst at a lower level of effort and expertise required by the user to predict voter turnout. Faced with cost constraints, we believe our model would be a viable alternative for a pollster or a campaign manager to predict voter turnout.

# References

Nickerson, David W. and Todd Rogers. *Political Campaigns and Big Data*. Journal of Economic Perspectives, Volume 28 Number 2, Spring 2014, pgs 51-74