MS&E 226 Project Part 3

Zhaoyu (Joe) Lou and Carina Zhang

December 2018

1 Test Set Predictions

1.1 Classification

We use the logistic model with 20 predictors from part 2 of the project. As we discussed in Part 2, we chose AUC as a measure of performance of our model because of the prevalence of one party (>0.8 Republican counties). Our test AUC is 0.964617, sensitivity is 0.7592593, specificity is 0.9747573, and accuracy is 0.9373997. Overall, our classifier has a better AUC, a better sensitivity, and a better specificity, meaning that it captures both the minorities in the data correct without sacrificing overall accuracy. The better test results might come from favorability of our test data due to randomness.

1.2 Regression

We use the XGBoost model with 100 decision trees, as we found in part 2 of the project. Surprisingly, we find that we get a test error (100 x RMSE) of 4.222; our regressor does slightly better on the test set than it does on cross validation! This means that our CV estimate of test error was slightly underestimating the true test error. We likely happened to end up with an easy test set for this task.

2 Inference

In the inference part, we kept the original training/test set split. We picked the OLS model as the model with 13 handpicked-covariates to study inference. We have previously tabulated best-performing models under Lasso, Ridge, and other regression models, and this 13-covariate model was the best performing OLS model. The covariates in this model as well as the OLS coefficients are tabulated in Appendix A.

2.1 Significance

For this model, we find that 10 out of the 13 covariates have statistically significant nonzero value at a significance level of 99.9%. Of the remaining covariates, one was significant at the 99% level and one at the 95% level. The coefficients and their p values are tabulated below. In particular, the covariates significant at the 99.9% level were: log(population), percent population under 18, percent female, percent hispanic, percent residing for over 1 year, percent high school graduates, percent homeowners, log(occupied house value), average family size, and log(income per capita).

Covariate	p-value
log(population)	$< 2 \times 10^{-16}$
percent under 5	1.95×10^{-3}
percent under 18	2.93×10^{-5}
percent over 65	0.0139
percent female	$< 2 \times 10^{-16}$
percent hispanic	8.60×10^{-6}
percent residing for over 1 year	5.79×10^{-6}
percent speak foreign language	0.228
percent high school graduates	$< 2 \times 10^{-16}$
percent homeowning	$< 2 \times 10^{-16}$
log(occupied house value)	$< 2 \times 10^{-16}$
average family size	6.65×10^{-5}
log(income per capita)	1.32×10^{-14}

Statistical significance in this case means that assuming the OLS estimate of the parameters is unbiased and normal, it is unlikely to have seen an extreme a value of the estimate we saw if the true parameter were zero.

2.2 Significance on Test Data

When we fit the model on the test set instead of the train set, the significances of some of the covariates change. In particular, percentage under 5, percentage over 65, percent, hispanic, and average family size are no longer significant even at the 95% level, and percent foreign language speaking becomes significant at the 99.9% level. The full table with p values is in Appendix B. Interestingly, comparing the p values from the model trained on the training set to these p values shows that in this model, the p value of every covariate except for percent residing for over 1 year and the percent speaking a foreign language go up, while the p value for the percent speaking a foreign language in particular goes down considerably. This suggests that perhaps in this dataset foreign language speaking and residency were correlated with the other covariates, leading to their being used to explain some of the variation that previously was explained by the other covariates.

It's worth noting here that the t-test results should be taken with a grain of salt, because while the distribution of voting percentages (displayed in Figure 1) looks fairly normal, the assumption of independent normal errors seems difficult to justify given possible geographic correlations, lack of any randomization to control for important but unaccounted for factors, and other lack of data on the potential distribution of errors.

It's also possible that some of our conclusions about significance were wrong, since we partook in multiple hypothesis testing (testing each coefficient individually). At a significance level of 0.99, we would have around a 13% chance of falsely concluding significance on at least one test (assuming all the null hypotheses were true). Indeed, applying the Bonferroni correction makes our coefficient on the percent population under 5 and the average family size no longer significant in the train data, which are two of the covariates which had much higher p values in the test set. A number of our remaining covariates are still highly statistically significant even under the Bonferroni correction, though, which leads us to be optimistic that we captured some important correlations.

Also note that the test set is much smaller than the train set (700 rows and 2300 rows, respectively), so we could be overfitting the test set (although this doesn't appear to be the case; the model fitted on the test set does acceptably well on the train set).

The biggest concern is post selection inference, which we will discuss in more detail in 2.4.

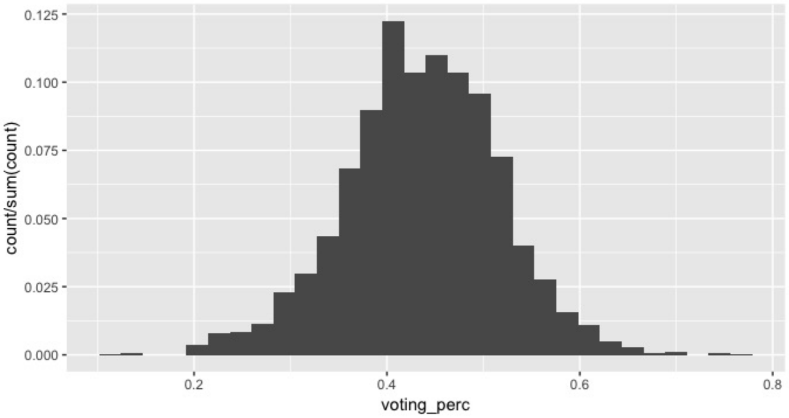

Figure 1: Distribution of voting percentages.

2.3 Bootstrap Estimation

We use the bootstrap with 3000 replications to estimate 99% confidence intervals for each coefficient using the bias corrected and accelerated confidence intervals in R. The results are tabulated below. All the confidence intervals match the hypothesis tests conducted earlier.

Covariate	Lower CI Bound	Upper CI Bound
log(population)	-0.0176	-0.0144
percent under 5	-0.011	-0.0005
percent under 18	-0.0057	-0.0008
percent over 65	-0.0002	0.0022
percent female	0.0045	0.0079
percent hispanic	-0.0015	-0.0003
percent residing for over 1 year	0.0005	0.0023
percent speak foreign language	-0.0003	0.0011
percent high school graduates	0.0017	0.0032
percent homeowning	0.0014	0.0025
log(occupied house value)	0.0214	0.0454
average family size	-0.0532	-0.0050
log(income per capita)	0.0436	0.0912

2.4 Comparison to Full Model

We compare the significances of the covariates in the model trained on handpicked covariates to an OLS model trained on all the covariates without data transformation or interactions. We find our conclusions about the significances of some of the covariates change as a result; in particular, the non log population is no longer significant, percent population over 65 becomes significant but percent population under 18 is no longer significant, percent foreign language speaking becomes significant, and average family size is no longer significant. Some of the covariates in the full model but not in the handpicked covariates are significant, including the percent of foreign born citizens, number of housing units, change in nonfarm employment, and percentage of multi-unit houses.

We are somewhat skeptical of these results, however; highly skewed covariate distributions with noticeable outliers likely contributed to lack of significance in some covariates (for example, the distribution of populations shown in Figure 2 has some extreme outliers but is much more normally distributed when we take a log, as we do in the handpicked covariates). Further, there are substantial correlations and collinearity in the set of all covariates. For example, we have covariates breaking down the population by race, but these percentages sum to 1, the intercept column. As another example, we have data on high school graduates and college graduates, but these covariates are likely highly correlated, as are covariates such as percent

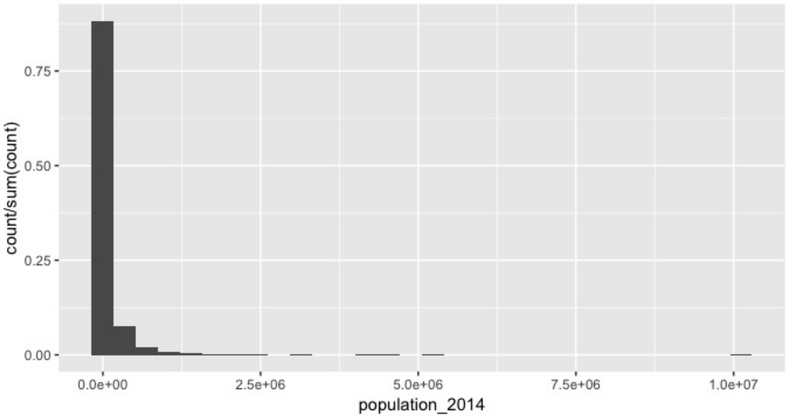

Figure 2: Distribution of populations.

hispanic and percent hispanic owned buisnesses. We use the VIF package in R to detect and assess the multicollinearity of our covariates. Generally, a threshold for collinear covariates is a VIF bigger than 10; when we have all covariates, we find that only 25 out of the 55 covariates do not exhibit a high degree of collinearity (indeed, some of them have extremely high VIF values higher than 100)! As discussed earlier, this is not too surprising given the nature of our data. These correlations make the direct application of the t-test inadvisable and likely contribute to our suddenly changed results in these covariates. In our handpicked covariate set, none of our covariates exhibit high collinearity, which increases our confidence in the original conclusions.

The biggest problem might lie in post-selection inference. We handpicked covariates based on looking at the correlations with the outcome in the training set and we assessed the significance of the coefficients using that same data. Thus we likely biased our p values downwards, which could go further in explaining why almost all our p values went up when we fitted the model on the test set.

2.5 Causal Inference

Under the observed effects model, we believe that while tempting, it is not advisable to treat the associations highlighted by this model as causal relationships. We believe that some of these covariates are proxies for other confounders which have the true causal explanation; for example, the percent of the population under 5 is probably a proxy for the percent of the population who are busy mothers who don't have time to vote. In such a case, causal inference would be invalid due to a selection bias - counties with lots of young children likely tend in the population to have lower voting rates. Simply opening a foster home or increasing the number of children in some other way probably wouldn't drastically change the voting rate. It would certainly be infeasible to randomly assign counties to have different demographic compositions, and within the dataset we had we failed to find any valid instrumental variables due to the difficulty in establishing the exclusion restriction.

3 Discussion

3.1 Practical Use of the Model

Practically, the voter turnout regression model would probably be used as a prelude to causal inference and public policy decisions. Since high voter turnout is quite important for the proper functioning of a democratic society, this model could be used to understand what affects voter turnout and what policy changes might be effective in increasing voter participation. However, since we highlighted earlier that causal inference may not be valid on this data, we believe that the results from this model are better interpreted as suggestions for further directions to explore - covariates deemed significant in this model may be worth conducting further study or experiments on to better assess the causal effect they have. Given these causal inferences, we then might be able to better understand why those covariates lead to higher or lower voter turnout, and then make public policy decisions to address those concerns and improve voter participation.

3.2 Long-term Applicability

In terms of the long term applicability of this model, it's worth noting that our data was from a single presidential election, so it's unclear how well the results will generalize to all voting at any time. However, demographics tend to change fairly slowly over time, so we believe that at least in the next few years the conclusions from the model will stay valid. The modeling paradigm could also easily be extended to incorporate more data on more and other types of elections, which would enhance its generalizability over time and over election types.

3.3 Notes for Future Users

One aspect of data cleaning that any user of this model would need to know is the exclusion of Alaskan data - these results may or may not hold there. The concerns raised earlier regarding multiple hypothesis testing and post selection inference leading to varying conclusions regarding some of the covariates would also be worth mentioning.
Another aspect on the inference part is that we did detect collinearity and some post-selection inference problems, which were discussed in the previous parts of this project.Thus, if anyone want to modify our model with other covariates, they need to be cautious when they interpret the coefficients of the variables since they might inadvertently introduce the collinearity that our model was chosen to avoid.

3.4 Improvements for Future Analyses

In retrospect, collecting additional data on election outcomes for other types of elections and other years might have enhanced the generalizability of our dataset. Our demographic data was quite through, so there we didn't feel that there were critical covariates missing from our demographic analysis; however, one interesting addition to our study would be to also have data on campaign advertisements and other efforts to increase voter turnout in the area. This would allow us to also assess the efficacy of different common methods for addressing the underlying problem of voter participation.
In terms of data exploration, we could have done more data visualization and analysis on a more detailed level to gain more insight towards possible conclusions. It would also have been interesting to combine results from the two models we built to gain deeper insights. For example:
1. What candidates within each party have results that are the most negatively correlated?
2. What candidates were correlated with higher voter turnout? What strategies did those candidates employ to increase voter turnout?
3. What insights can we discover from geographically mapping the data?

In terms of prediction and inference, we could have increased the significance of our conclusions by doing covariate selection on a separate held out set of data to avoid the post-selection inference issues we faced in this part as well as the bias issues from not fully cross validating the selection. This was made difficult in our dataset due to the relatively small number of examples, but if we included results from other years/elections we would be able to have enough data to make this split within the training set.
It would also have been interesting to do a logistic regression on the same voter turnout task; since we were predicting the proportion of voters and that proportion is limited to the range [0,1], the output of the logistic regression has a natural interpretation as the proportion. This would also avoid potential extrapolation issues, though perhaps at the cost of interpretability; the coefficients on the logistic model would certainly not have as direct an interpretation as the linear model (log odds don't exactly lend themselves to intuitive explanation).

4 Appendices

A Regression Coefficients

Covariate	Coefficient
log(population)	-0.0145
percent under 5	-0.0058
percent under 18	-0.0032
percent over 65	0.0010
percent female	0.0063
percent hispanic	-0.0009
percent residing for over 1 year	0.0014
percent speak foreign language	0.0003
percent high school graduates	0.0025
percent homeowning	0.0019
log(occupied house value)	0.0339
average family size	-0.0297
log(income per capita)	0.0662

B Regression p values on Test Data

Covariate	p-value
log(population)	4.87×10^{-15}
percent under 5	0.302
percent under 18	4.5×10^{-4}
percent over 65	0.0582
percent female	4.49×10^{-5}
percent hispanic	0.276
percent residing for over 1 year	9.55×10^{-7}
percent speak foreign language	2.47×10^{-4}
percent high school graduates	3.11×10^{-7}
percent homeowning	0.00282
log(occupied house value)	3.08×10^{-7}
average family size	0.719
log(income per capita)	2.82×10^{-6}