North Carolina in the 2016 Presidential Election

Ian Gardner & Jake Berberian

Overview

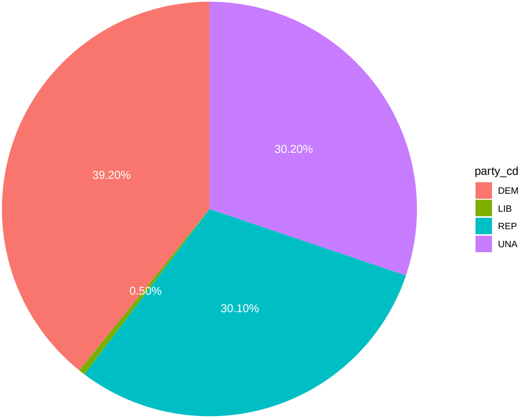
NC Overview

- North Carolina and its 100 counties hold 15 electoral college votes.
- Give demographic statistics?

Data

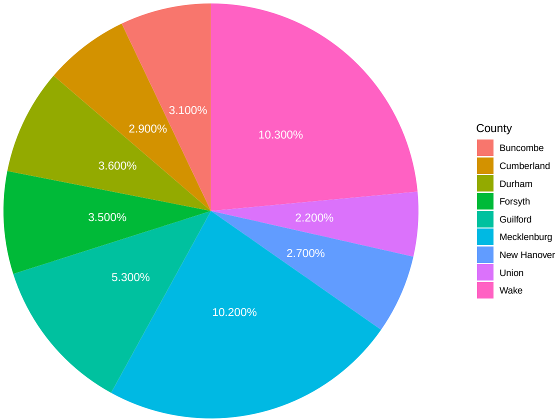
County	stats_type	vtd_abbrv	party_cd	race_code	ethnic_code	sex_code	age
Alamance	voter	169	UNA	O	NL	F	Age 26 - 40
Alamance	voter	263	REP	W	UN	M	Age 26 - 40
Alamance	voter	263	REP	W	UN	M	Age 26 - 40
Alamance	voter	263	REP	W	UN	M	Age 26 - 40
Alamance	voter	263	REP	W	UN	M	Age 26 - 40

By party affiliation

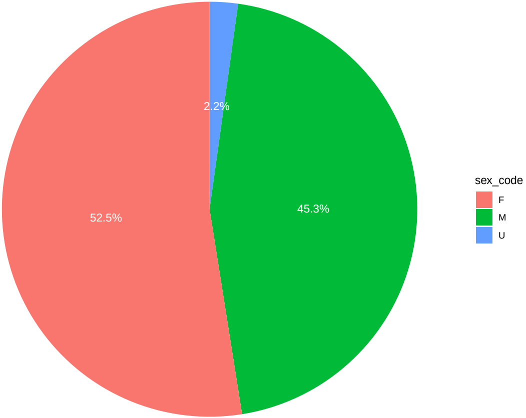


By county

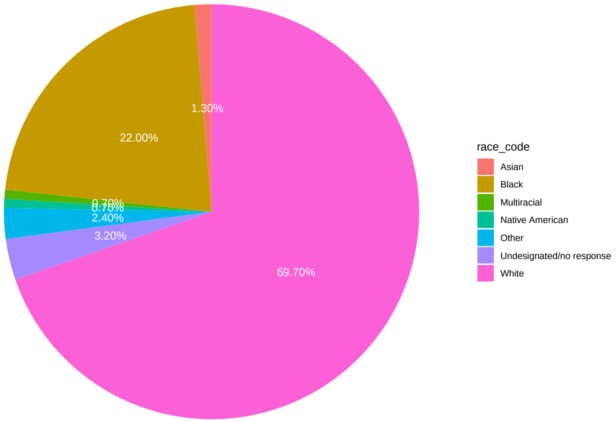
- Below are some of the larger counties in NC, by voter turnout.
  - Any county that represents over 2% of the data is below.



By sex

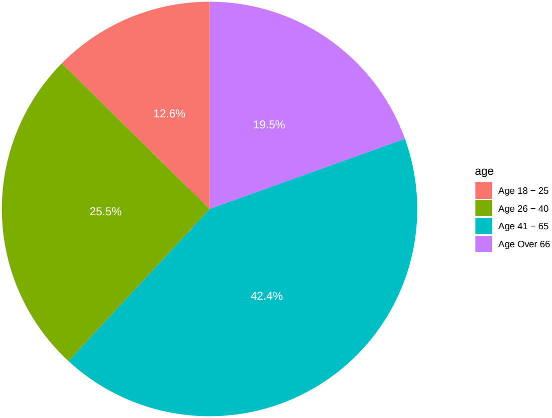


By race



1.30%
22.00%
0.70%
0.70%
2.40%
3.20%
69.70%
race_code
Asian
Black
Multiracial
Native American
Other
Undesignated/no response
White

By age



Multinominal Logistic Regression

An Overview

- Models how multinomial response variable Y depends on a set of k explanatory variables $X = (X_1, X_2, ..., X_k)$.
  - Is classified as a generalized linear model where the random component assumes that Y ~ Multinomial(n, π).
  - π is a probability success vector for each given Y category.
- The link function is generalized logit.
  - A link function transforms the probabilities of a categorical variable into a continuous, unbounded scale.
- Since our data is nominal, we must perform nominal regression
  - Nominal = unordered
- PMF: $\frac{n!}{x_1!,...,x_k!} p_1^{x_1} ... p_k^{x_k}$

Considerations

- Our sample size should be large enough, as multinomial regression uses maximum likelihood estimates. With well over 800,000 observations, our data satisfies this assumption.
- Separation between outcome and predictor variables
- No NAs
    - All NA observations (0 in here) have been dropped from this dataset.

Applications in R

- To simplify computation, we'll look at strictly Orange County data.
- We'll split this data into training and testing data, through random sampling.
  - This allows for cross-validation, or checking the accuracy of our model.

- Using the `nnet` package in R, we'll use the function `multinom()` to carry out our multinomial logistic regression.
  - Democrat is our baseline level of *party_cd* when the regression is run.
  - Coefficients of our regression are outputed below.

	(Intercept)	race_codeB	race_codeI	race_codeM	race_codeO	race_codeU	race_codeW	sex_codeM	sex_codeU	ageAge 26 - 40	ageAge 41 - 65	ageAge Over 66
LIB	-3.7861832	-2.3064401	1.5611841	-0.5995370	-0.5383042	1.2909854	0.1959223	1.120241	0.3272172	-0.2947151	-1.8337206	-2.6275935
REP	-3.7835548	-0.5651973	1.5202473	1.4654979	1.6461137	1.2772980	2.0368289	1.244312	1.7593633	0.6095434	0.0988897	-0.5933063
UNA	-0.5086557	-0.9988455	-0.3132486	-0.3736255	0.3662906	0.6834832	-0.1303131	1.156793	1.9507635	0.0804066	0.0477630	-1.1916412

Importance of variables

- Unlike `summary()` with linear and generalized linear regression models, `multinom()` doesn't output the importance of each variable.
    - For small p-values, let's say $\alpha = 0.05$, we'll consider the variable to be "important."
    - No variables that are entirely insignificant, so we'll keep all of them in there.

	(Intercept)	race_codeB	race_codeI	race_codeM	race_codeO	race_codeU	race_codeW	sex_codeM	sex_codeU	ageAge 26 - 40	ageAge 41 - 65	ageAge Over 66
LIB	0.00e+00	0.0406755	0.1873359	0.5966291	0.6342024	0.0919933	0.7140131	3.78e-05	0.7187970	0.3230317	0.0000032	0.0003778
REP	0.00e+00	0.2081655	0.0788413	0.0022676	0.0002086	0.0067968	0.0000000	0.00e+00	0.0000002	0.0000001	0.3704551	0.0000463
UNA	1.85e-05	0.0000000	0.5468666	0.1165941	0.0533488	0.0005840	0.2678829	0.00e+00	0.0000000	0.3120266	0.5161988	0.0000000

Cross-validation

	DEM	LIB	REP	UNA
DEM	5729	49	936	2440
LIB	0	0	0	0
REP	0	0	0	0
UNA	1546	64	1048	3379

- The above table is our confusion matrix.
  - Horizontal is our testing data results and vertically is our predicted results. Diagonal is correctly classified
- Interesting that our model *never* classifies a voter as a Republican or Libertarian.
  - Possibly due to differing demographics amoungst those parties
- Our classification rate is 59.96%.

Intepretation of Results

	race_codeB
LIB	-2.3064401
REP	-0.5651973
UNA	-0.9988455

A one-unit decrease in the variable *race_codeB* is associated with the decrease in the log odds of being a Libertarian vs. a Democrat in the amount of 2.31.

Risk Ratio

- We can find the corresponding risk ratios, by exponentiation of all of our log odds.
- The relative risk ratio for a one unit increase in the variable *race_codeB* is 0.0996 for being a Libertarian vs. a Democrat.

	race_codeB
LIB	0.0996152
REP	0.5682480
UNA	0.3683044

Discriminant Analysis

Overview

- Clustering technique that is closely related to PCA.
- Model the distribution of predictors X separately (opposed to logistic regression) in each of the response classes and use Bayes' theorem to flip these into estimates for $\Pr(Y = k|X = x)$
  - Bayes' Theorem: $\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A)+\Pr(B|\neg A)\Pr(\neg A)}$
- Use it when classes are well-separated, if n is small & Y is approximately normal, and is popular if there are more than two response classes.
- LDA vs QDA
  - LDA attempts to create a linear boundary between classifiers, while QDA creates a non-linear boundary.

Applications in R: LDA

```r
library(MASS)
nc_lda <- lda(party_cd ~ race_code + sex_code + age,
              data = train_oc)
lda_pred <- predict(nc_lda, test_oc)
```

- The below output is only the first few terms of the discriminant analysis.

	LD1	LD2	LD3
race_codeB	-1.1155547	-0.7498365	-0.6030964
race_codeI	-0.0775873	-1.7851277	5.5858654
race_codeM	-0.2535873	-1.6064455	-0.8746791
race_codeO	0.5433923	-0.6321081	-1.1507063

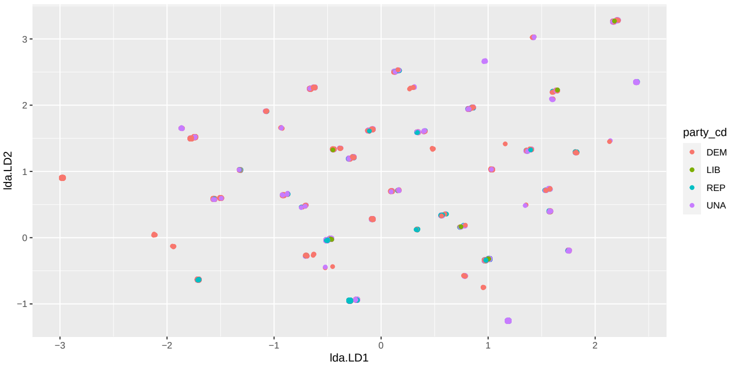
Cross-validation: LDA

- Again, we see that our model fails to predict any observation to be Republican.
- We see our classification rate on our testing set is just barely smaller than that of our multinomial logistic regression.

	DEM	LIB	REP	UNA
DEM	5718	49	932	2431
LIB	8	2	2	8
REP	0	0	0	0
UNA	1549	62	1050	3380

Classification Rate
0.5990389

- We see that linear discriminant analysis does not do a great job either. We would like to see four distinct clusters, one for each party.
- It's safe to conclude that LDA does not perform well and our classification rate is probably misleading and higher than it should be.
- There are a ton of overlapping points here.



Applications in R: QDA

```
nc_qda <- qda(party_cd ~ race_code + sex_code + age,
                data = train_oc)
qda_pred <- predict(nc_qda, test_oc)
```

- The below output are the prior probabilities necessary for Bayes' Theorem. These are the proportion of training observations from each group.
  - For example, there are approximately 48% of the training observations in the Democrat group.
  - As expected, these sum to 1.

	Prior Probabilites
DEM	0.4792186
LIB	0.0079979
REP	0.1313754
UNA	0.3814082

Cross-validation: QDA

- We see that QDA predicts some Republicans, but is still somewhat concentrated in it's prediction. This is to be expected.
- However, we see a much lower classification rate. This is probably more accurate too. Less overfitting of the *testing* data.

	DEM	LIB	REP	UNA
DEM	2548	8	230	858
LIB	79	5	16	119
REP	4120	88	1630	3912
UNA	528	12	108	930

Classification Rate
0.3365809

Conclusions

Conclusions

- One thing that hasn't been discussed is that the party code UNA indicates unaffiliated. This could throw off our analyses greatly, as unaffiliated voters generally do not take a certain demographics like the two major parties do.
  - One of the fastest growing electorates.
- Neither of our predictive techniques performed that well on our data.
  - Could be due to the nature of only looking at one county.
  - This is why we poll. If it were this easy, then elections would be no fun.
- If we had to pick a model, we'd likely go with our multinomial logistic regression.
  - Discriminant analysis usage may not be the best, with a sufficiently large n and some colinearity.

Sources

Sources

- https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/
- James, Gareth, et al. *An Introduction to Statistical Learning with Applications in R.* 7th ed., Springer, 2017.