Voting Behavior

An Explanatory Model

Jiying Zou

Introduction

Political candidates receive support from both those supporting their visions and from those identifying with them culturally. To what extent will a voter's race, income level, and geographical location influence their vote? Will Hispanics support a Black candidate similarly as minority groups do? This project builds and evaluates an explanatory model for voting behavior in the 1988 Democratic presidential primary election. Candidates include a Black minister, Jesse Jackson, and three White candidates. Results show that Black voters have 31 times higher odds of voting for Jackson, while Hispanics support Jackson to a lesser degree. Furthermore, support for Jackson varies over precincts and income levels.

Data Description

The dataset used comes from an exit poll held by the Field Research Corporation, a private research firm performing consulting work for government interests. The poll surveyed 1867 voters on race, income level, precinct of residence, and candidate supported. Five racial groups, coded 1-5, are White, Hispanic, Black, Asian, and other. Eight income groups are coded 1-8 and represent \$0-10k, \$10-20k, \$20-30k, \$30-40k, \$40-50k, \$50-60k, \$60-70k, and \$70k+ annual income. Thirty-nine precincts are represented. Since sampling methods are unknown, the data is treated as representative of the entire small city population.

Zeroes in the dataset are replaced with N/A's, representing missing data. Only race and income variables had missing values. The eight individuals missing both values are removed due to lack of information. Remaining missing values are distributed reasonably evenly throughout race and income groups and represent <10% of all cases, so removal should not skew results noticeably.

Each variable is treated categorically, with numerical values representing individual groups. Figure 1 shows that the percentage of each income group supporting the Black candidate decreases as we move up in annual income categories. Different races support Jackson to differend degrees, and these proportions vary by precinct (Figure 2). In fact, some precincts have data missing on certain races' voting preferences.

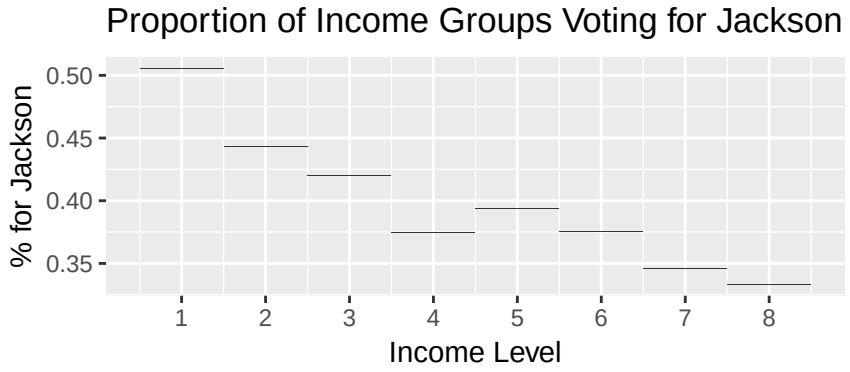

Figure 1. Over half of those with annual income \$0-10k voted for Jackson, and this percentage generally decreases as annual income increases. (Annual income group coding: 1 = \$0-10k, 2 = \$10-20k, 3 = \$20-30k, 4 = \$30-40k, 5 = \$40-50k, 6 = \$50-60k, 7 = \$60-70k, and 8 = \$70k+ annual income)

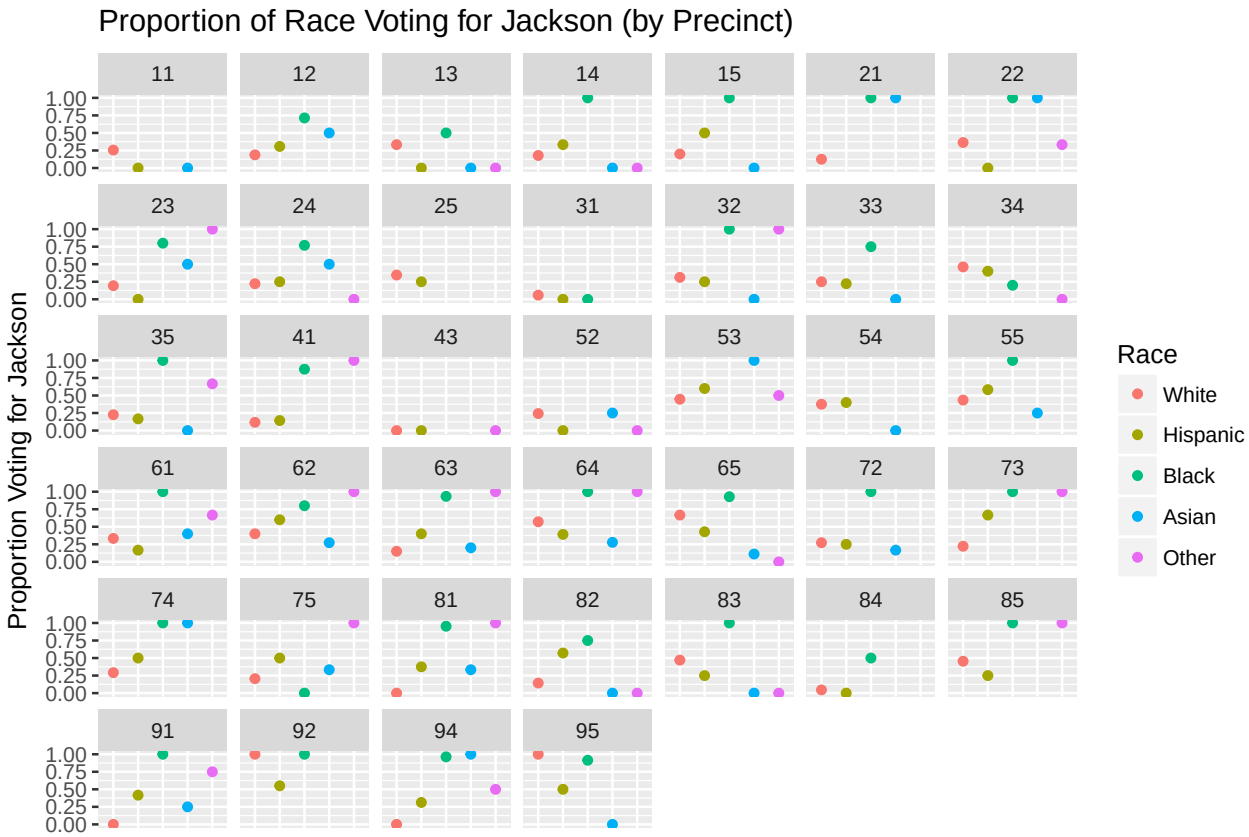

Figure 2. The percentage of each racial group voting for Black candidate Jackson shows some variation across different precincts. White voters consistently tend not to vote for Jackson, while most Black voters do. Some precincts, such as 11, 25, 31, 43, and 84, are missing data on some races. Note: low numbers of individuals present in some racial groups in precincts may skew proportions.

The mosaic plot explores income distribution within and between races (Figure 3). White voters make up the sample's largest gorup; those sampled in this group have a rather uniform annual income distribution. The second largest group is Black, whose annual incomes tend to fall in the lower ranges ($0 to $30k per year). Other minority groups also are underrepresented amongst higher incomes.

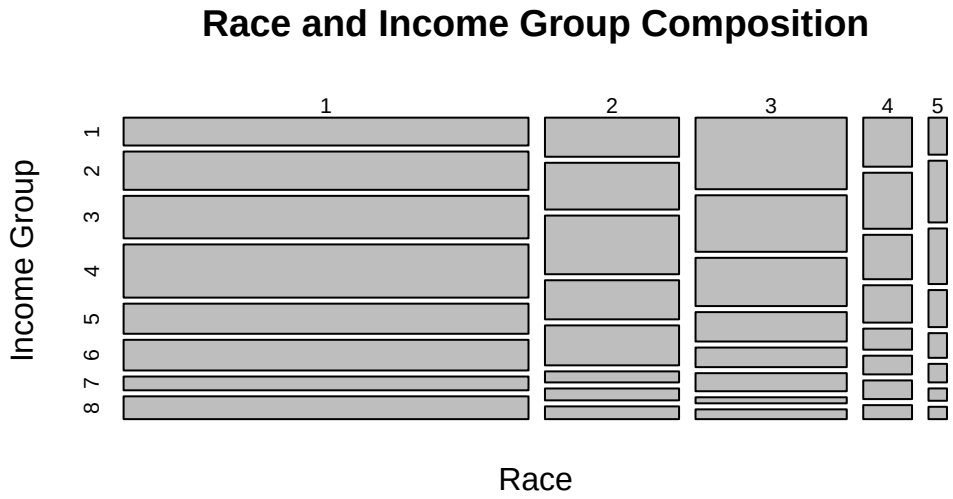

Figure 3. Visualizing imbalances in the sample's racial and income group compositions. Black individuals (3) are overrepresented in the lowest three income groups, which indicate $0-30k annual income. There are very few Asian (4) and other (5) race individuals in the sample. (Race groups coded: 1 = White, 2 = Hispanic, 3 = Black, 4 = Asian, 5 = Other)

Visualizing the relationship between race, income, and chance of voting for Jackson illustrates that black individuals across all income groups tend to vote for the black candidate, while income groups individually have almost an even split the votes The two variables do not seem to interact very much (Figure 4).

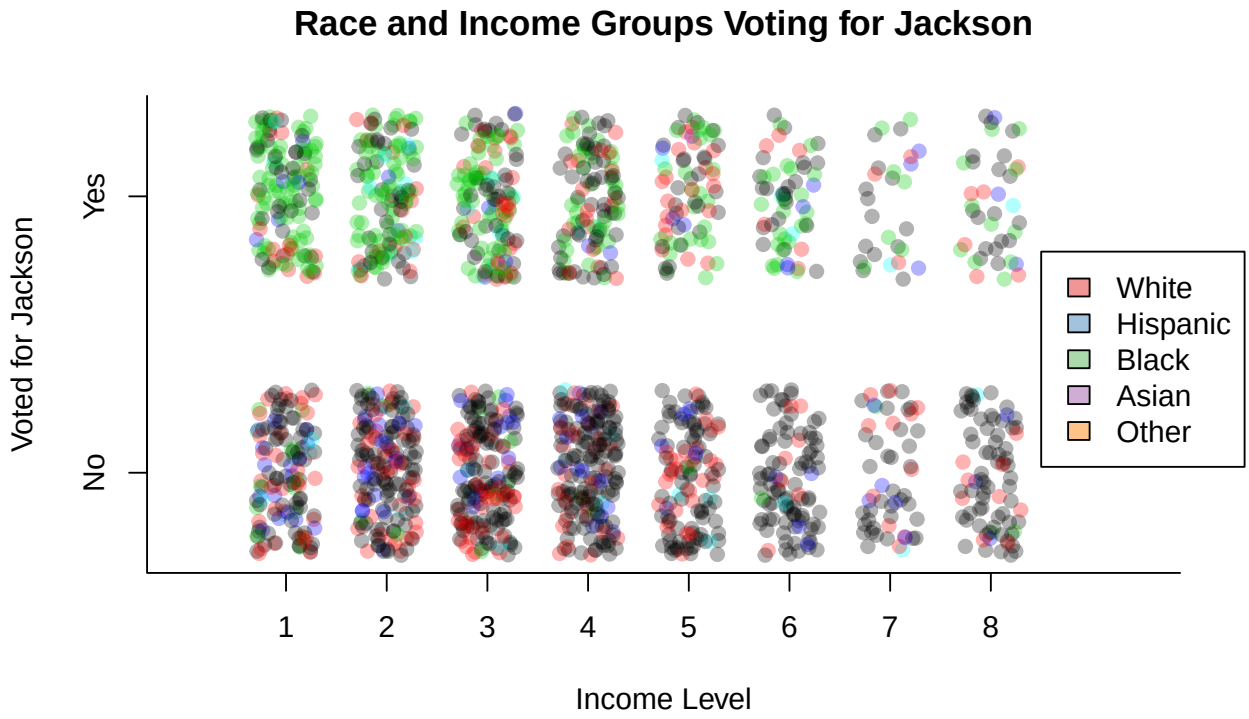

Figure 4. Lower income groups and Black individuals exhibit a larger chance of voting for Black candidate Jackson. The plot is jittered about income levels and responses to avoid overplotting. (Annual income group coding: 1 = \$0-10k, 2 = \$10-20k, 3 = \$20-30k, 4 = \$30-40k, 5 = \$40-50k, 6 = \$50-60k, 7 = \$60-70k, and 8 = \$70k+ annual income)

Main Results

I first fit univariate and multivariate models with and without interactions to quantitatively explore voting behavior relationships. Results show that non-White races (besides Asians) have a greater odds of voting for Jackson, with the odds increasing by $e^{3.29} = 26.84$ times alone amongst Black voters! This corresponds to a $(\frac{p}{1-p}) = 26.84 \longrightarrow p \approx 0.9640 = 96.4\%$ chance of voting for Jackson amongst Black voters. Certain precincts have anywhere from about 7 to 190% higher odds of voting for Jackson. Income groups also show significance, but much of this disappears in the presence of other variables. Few significant interactions are detected.

Four unique models based on previously significant characteristics are chosen by stepwise iterations, Mallows' Cp criteria, adjusted R^2 criteria, and binomial LASSO regression. The binomial LASSO regression model considered interactions between race and precinct, which held a few significant terms from previous multivariate fits, but these interactions were not considered in the other models due to collinearity considerations. The models chosen by Mallows' Cp and adjusted R^2 criteria are ones exhibiting lowest Cp or highest adjusted R^2 value, respectively, within an exhaustive search for the top models of each size of up to 20 variables. Removing high leverage points during the search influences model selection for Mallows' Cp more than for adjusted R^2, narrowing down the former model size from 14 to 12 variables and barely altering the latter's top choice (Figures 5-8).

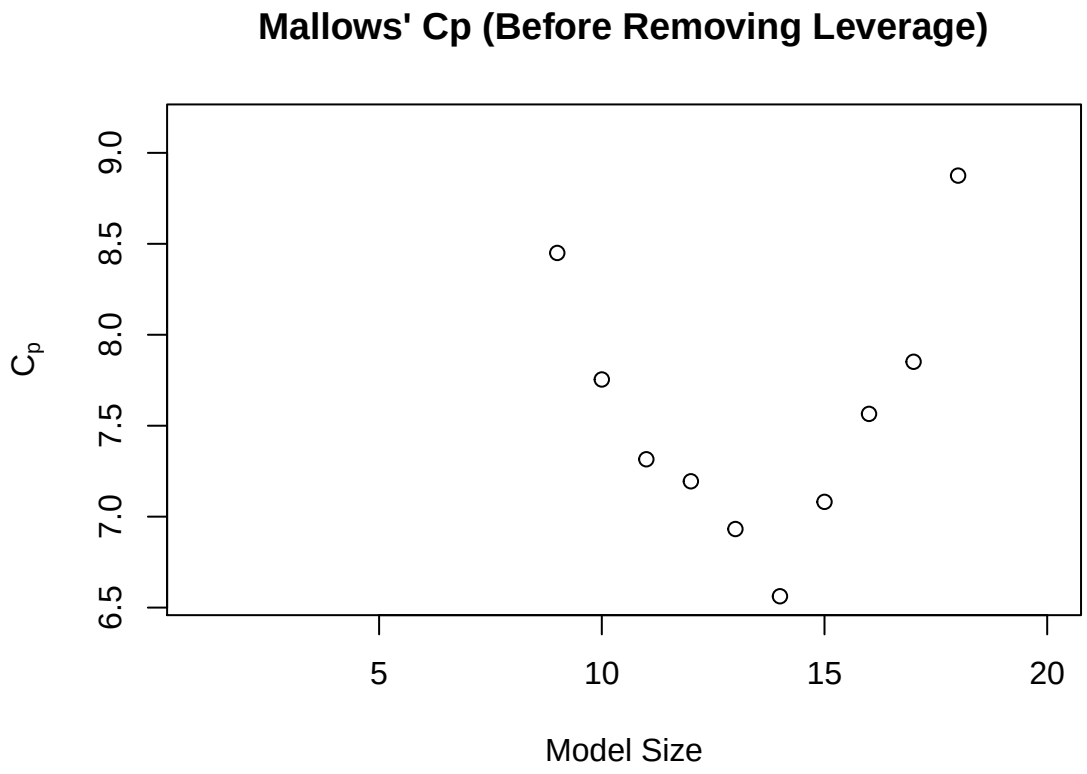

Figure 5. Model size chosen by Mallows' Cp criterion before removing high leverage points.

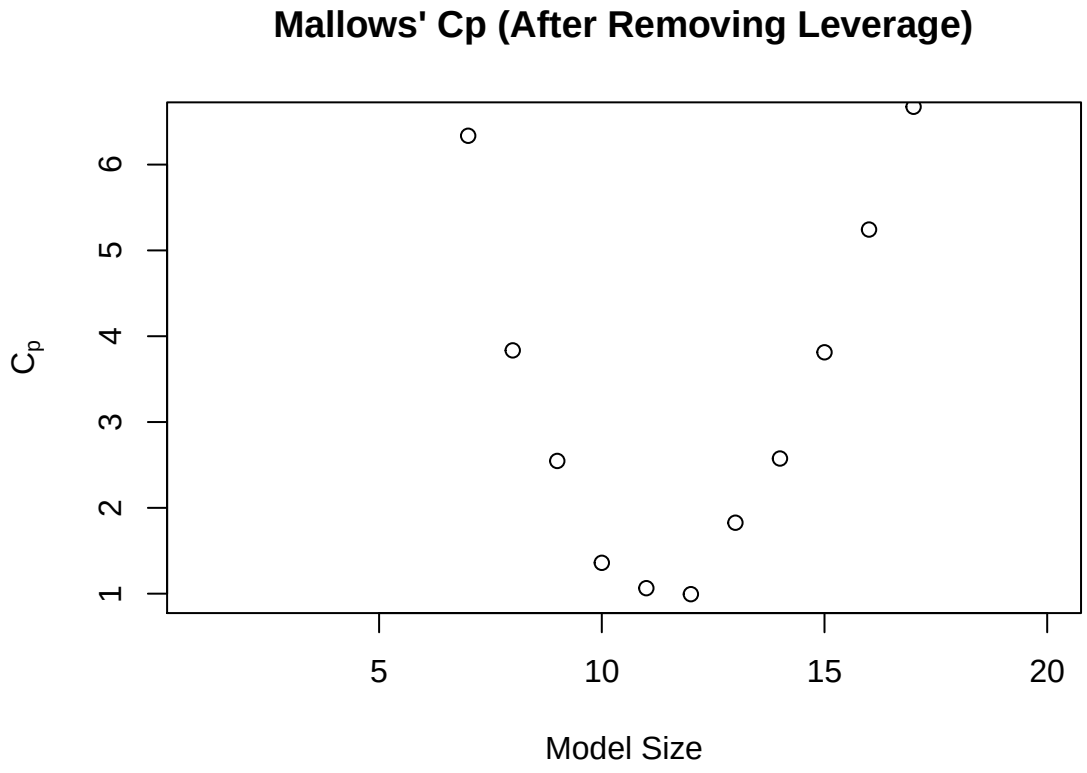

Figure 6. Model size chosen by Mallows' Cp criterion after removing high leverage points. Leverage points change model selection here.

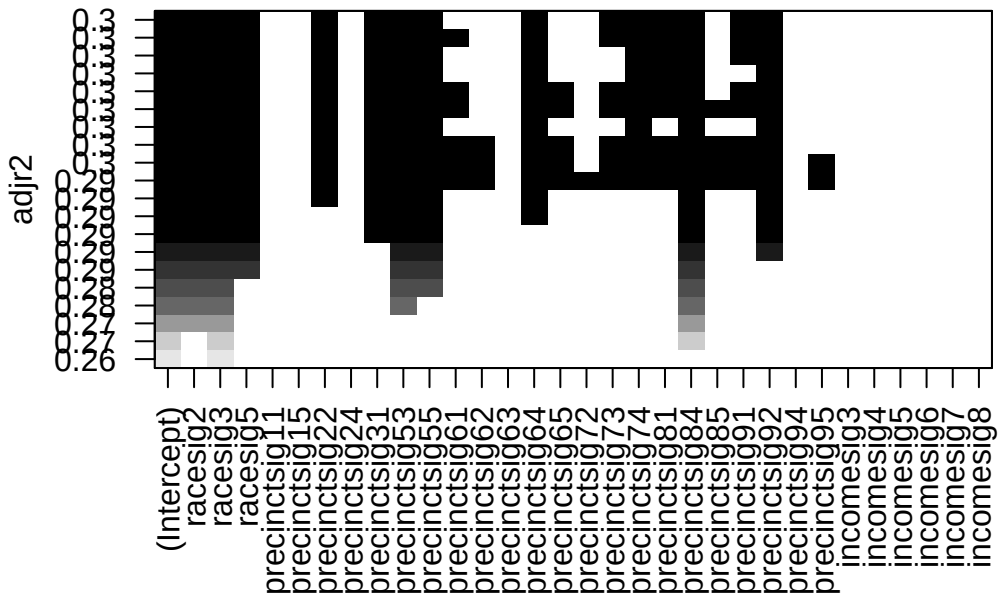

Figure 7. Models chosen by adjusted R^2 criterion before removing high leverage points.

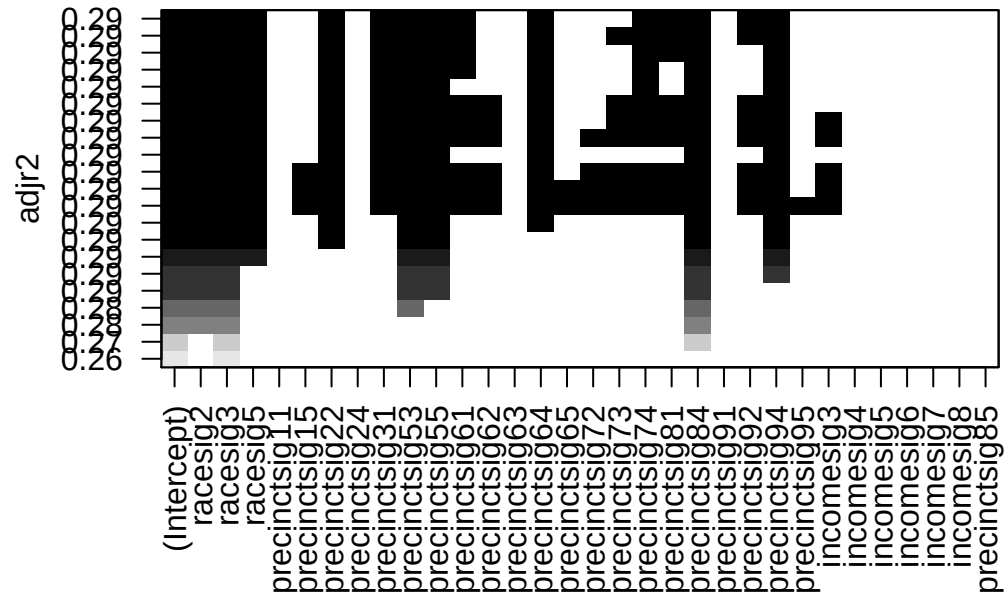

Figure 8. Models chosen by adjusted R^2 criterion after removing high leverage points. Leverage points barely change final model selection in this case.

This final model selected is the one found using adjusted R^2 criterion on an exhaustive search of the best model of each size up to 20 variables, is the best of the four, boasting a good tradeoff between interpretability and explanatory power (Table 1). It also has a comparatively decent adjusted R^2 of 0.2361, meaning this model explains away about 23.61% of data variation.

```
##                   Method Size Adjusted.R2    AIC
## 1              Stepwise   28      0.2589 1789.6
## 2           Cp Criteria   12      0.2361 1810.8
## 3 Adjusted R^2 Criteria   10      0.2344 1810.8
## 4        Binomial LASSO   36      0.2581 1807.4
```

Table 1. Four model selection methods, model sizes, and adjusted R^2 and AIC values of fit to entire dataset. Stepwise and binomial LASSO models are larger and less interpretable, but explain more variance. Cp and adjusted R^2 methods yield smaller, more interpretable models, yet perform almost as well as the others do.

The final model contains 12 variables, including 3 racial groups, 9 precincts, and no interactions. Coefficient values (log odds ratios) are displayed in the table below with their respective standard errors. Odds ratios are found by exponentiating the coefficients (Table 2).

```
##             LogOR   SE    OR P.value
## (Intercept) -1.17 0.08    NA  0.0000
## Hispanic     0.43 0.15  1.53  0.0042
## Black        3.43 0.21 30.77  0.0000
## Other        1.23 0.32  3.42  0.0001
## Precinct 22  0.52 0.30  1.68  0.0885
## Precinct 31 -2.42 1.14  0.09  0.0340
## Precinct 53  0.98 0.28  2.66  0.0005
## Precinct 55  0.90 0.27  2.45  0.0007
## Precinct 64  0.56 0.29  1.75  0.0547
## Precinct 74  0.53 0.36  1.70  0.1441
## Precinct 81  0.53 0.37  1.70  0.1495
## Precinct 84 -2.37 0.58  0.09  0.0000
## Precinct 92  1.22 0.44  3.39  0.0053
```

Table 2. Final model coefficients (log odds ratio), SE, odds ratio, and significance values. See analysis for interpretations.

Discussion

The final model asumes no sampling biases or recording errors, that the sample contains the entire population, and that individual responses are independent. However, these assumptions are overly ideal. For example, similar-minded voters from the same family violates the independence assumption. Given these limitations, this model is not highly generalizable to non-ideal scenarios. On the other hand, it reasonable that race affects voting behavior, and disparities between precincts may manifest from cultural differences and candidate campaigning efforts, so the model retains some validity.

Interpretation-wise, the coefficients are the log odds ratios associated with being versus not being in a certain variable category, holding other variables constant. The OR for the intercept is removed because it does not make sense to evaluate the odds ratio for a unit increase in intercept. Black voters have a 31 times increased odds of voting for Jackson compared to non-Black voters, an effect far surpassing Hispanic or other race's impacts (1.53 times and 3.42 times, respectively). Hispanic voters, controlling for precincts, do not seem to support Jackson as much as other minority voters do. Precincts effects are much milder; living in precincts 53 or 55 increases one's odds by about 2.5 times, while precinct 31 and 84 residents have a dramatically reduced odds (OR = 0.09). Figure 2 confirms this observation, where two races are missing and very few voters opted for Jackson. I hypothesize that income variables did not appear in the final model because its effect is sufficiently explained by race. Both minorities and lower income groups exhibited higher support for Jackson.

Regarding other models, the forward and both-direction stepwise (which yielded the same model) and LASSO models have slightly higher adjusted R^2 values and lower AIC than both Cp and adjusted R^2 models, but their composition is also much harder to interpret. This negligible 2% increase in explanatory power is not worth the extra complexity. Mallows' Cp and adjusted R^2 for `regsubsets()` exhaustive search yielded similar models with little disparity in final explanatory power or AIC and similar interpretability. Thus, the choice of the adjusted R^2 model is slightly arbitrary. There is little good reason to not prefer the Cp model, with one less variable and slightly only reduced explanatory power.

Summary

According to data analysis on an exit poll from the 1988 Democratic presidential nominee election, non-Asian minority voters and voters from certain precincts are more likely to support the Black candidate others are. Black voters are the most supportive, while Hispanic voters are supportive to a lesser degree than other minorities are. These results come from a model built based on adjusted R^2 criterion and explains about 25% of variation in the data.

References

Internet sources:

- For plotting help:
- http://stackoverflow.com/questions/3932038/plot-a-legend-outside-of-the-plotting-area-in-base-graphics
- https://www.datacamp.com/community/tutorials/15-questions-about-r-plots#q3)
- For help with creating a model matrix containing interactions: http://stackoverflow.com/questions/22649536/model-matrix-with-all-pairwise-interactions-between-columns)

Packages used:

- `DataComputing` – for various methods: http://data-computing.org/accessing-data-computing-data-and-software/
- `leaps` – for `leaps()` and `regsubsets()` exhaustive search: https://cran.r-project.org/web/packages/leaps/leaps.pdf
- `cvTools` – for cross-validation at the end: https://cran.r-project.org/web/packages/cvTools/cvTools.pdf
- `glmnet` – for binomial LASSO regression and cross-validation: ftp://debian.ustc.edu.cn/CRAN/web/packages/glmnet/glmnet.pdf
- `RColorBrewer` – for logistic scatterplot colors: https://cran.r-project.org/web/packages/RColorBrewer/RColorBrewer.pdf

Lecture slides (from Professor Deborah Nolan) consulted:

- LogisticRegression.html – for the mosaic plot and jittered scatterplot ideas and code
- ModelSelection.html – for the decision to zoom into Cp plots, and for `leaps()`, `regsubsets()`, and stepwise selection code

Lab examples (from Omid Solari) referenced:

- Lab 0412 – for stepwise code and cautions in model interpretation
- Lab 0419 – for LASSO graph code

Notes taken in class:

- On the topics of collinearity, PCA, cross-validation, model selection, AIC/BIC, ridge and LASSO regression, logistic regression, and generalized linear models

Note: Some of this code was taken from my own implementation of HW5 about baseball data analysis.

A huge thank you to Professor Nolan and Omid Solari for helping me debug throughout the project!