Stat 151a Final Project: Voting Behavior

Rushil Sheth

May 2nd 2017

Introduction

In this project we aim to see the association between race, income, and precinct to voting behavior. “In particular, do Hispanics who live in a precinct support the black candidate to the same degree as all other ethnic groups in that precinct? That is, consider whether voters in a precinct vote similarly regardless of race or income.” - Nolan.

We conclude that race and income play the biggest role in voting behavior. A description of the data we used to answer this question, followed by Exploratory Data Analysis, and the methodology and reasoning, ending with a discussion and summary will be given below.

Data Description

We use the data collected from an exit poll, conducted by Field Research Corporation, for the 1988 Democratic presidential primary. The exit poll is survey data collected from 1867 voters as they leave the polls. The information collected during the survey consists of race, income, educate and who they voted for. There is missing data for some of the race and income variables, so we remove these rows.

- **Precinct:** based on location of poll, identification number for precinct
- **Candidate:** Vote cast for 1 = Jackson; 2 = LaRouche; 3 = Dukakis; 4 = Gore; 5=Simon
- **Race:** Voter’s race 0 = missing; 1 = White; 2 = Hispanic; 3 = Black; 4 = Asian; 5 = other
- **Income:** Voter’s annual income (thousands of $) 0 = missing; 1 = 00–10; 2 = 10–20; 3 = 20–30; 4 = 30–40; 5 = 40–50; 6 = 50–60; 7 = 60–70; 8 = 70+

The voting behavior we are studying here is based on the race of the candidate. In 1988, Jesse Jackson, a Black minister, ran as a candidate for the Democratic nomination. The rest of the candidates were white. For this reason, we define our response variable:

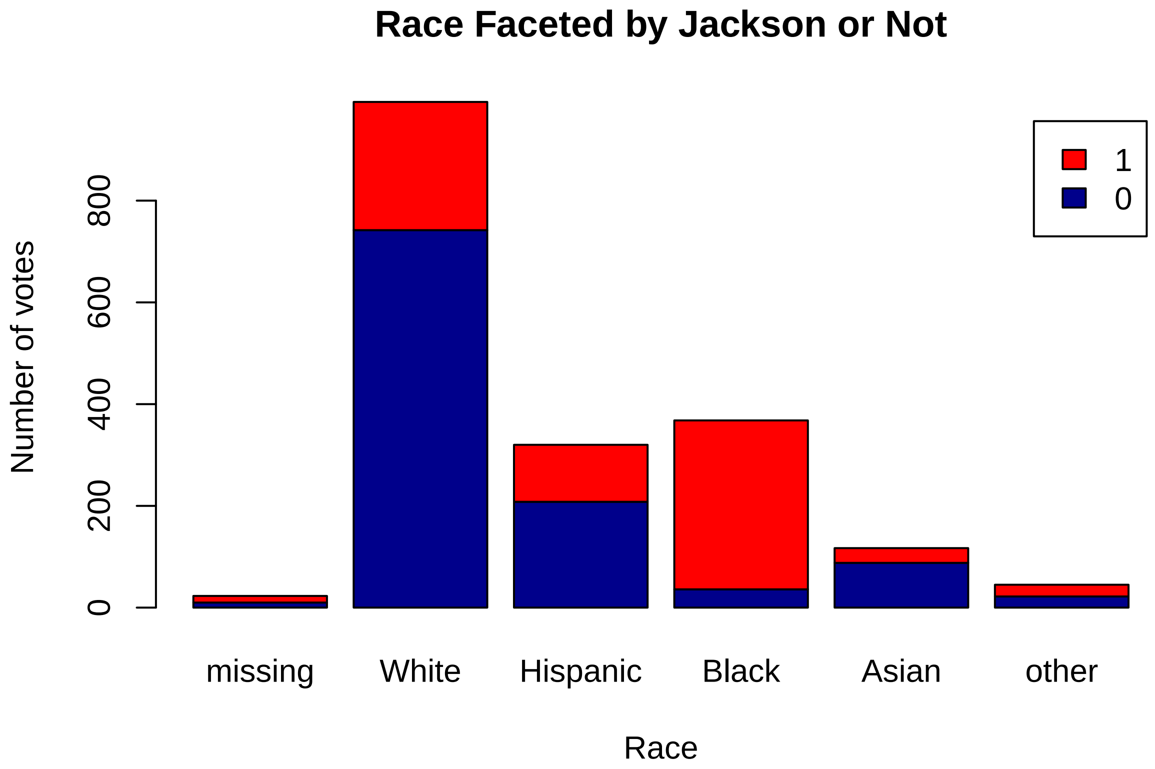
- **1 represents a vote for Jackson**
- **0 represents a vote not for Jackson**

We will now perform some EDA, while keeping this definition in mind about our response variable.

Exploratory Data Analysis

During this exploratory Data analysis we will look at the effect of the predictors(income, race, and precinct) and how they potentially explain the variability in our response variable. We will also look at potential interactions between income, race and neighborhood in regard to the response variable.

First, we look at the distribution of race and facet the race by amount who vote for Jackson(1) and individuals who did not vote for Jackson(0).



We can see that for some ethnicity there is a clear bias for Jackson, while others there is bias against Jackson. In particular, Blacks heavily favor Jackson, while Whites do not.

Now we will look more closely, at the precincts and their race distribution. This is hard to visualize graphically, so we will use a table instead.

```
11  12  13  14  15  21  22  23  24  25  31  32  33  34  35  41  43  52  53  54  55  61  62
```

0 0 0 2 0 0 0 1 2 1 0 0 0 0 0 0 1 0 0 1 1 0 0 0 1 50 32 41 30 35 17 44 32 27 29 18 51 56 14 34 18 40 38 50 20 47 3 5 2 3 13 4 3 13 0 2 4 8 4 2 8 9 5 6 8 1 2 5 5 13 6 5 3 0 14 2 3 11 1 4 5 13 0 1 1 4 5 6 9 0 0 0 0 2 22 16 4 1 2 6 1 7 1 1 4 4 0 0 1 1 0 3 0 0 4 1 1 4 6 12 5 1 0 1 1 0 0 3 1 2 0 0 1 0 1 3 1 1 1 2 0 0 3 1

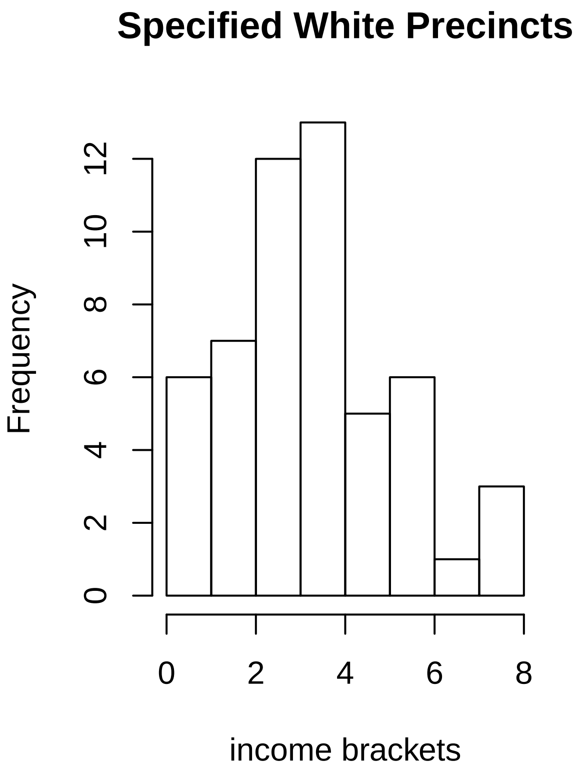
```
63  64  65  72  73  74  75  81  82  83  84  85  91  92  94  95
```

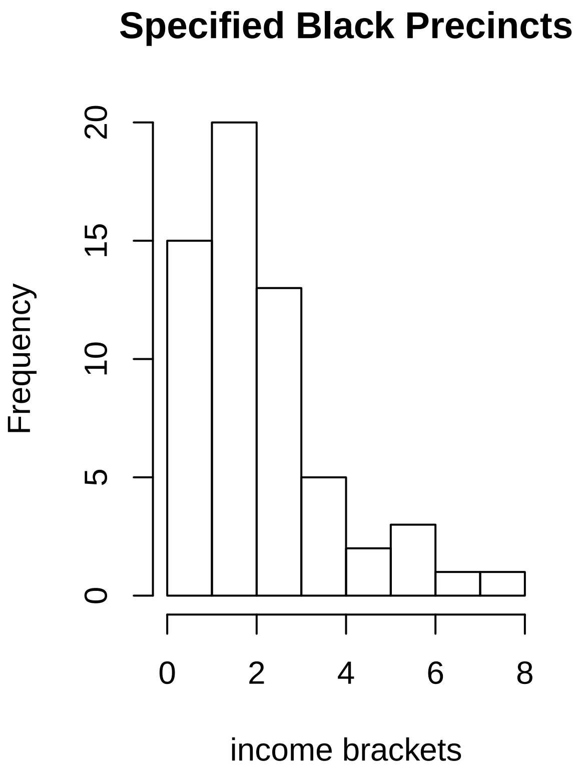
0 1 1 2 0 2 0 0 1 3 0 0 0 1 0 3 0 1 22 8 3 23 27 24 35 1 37 18 23 12 12 3 12 3 2 6 23 9 4 6 8 9 16 16 4 14 4 25 22 19 6 3 16 19 15 9 4 4 1 44 4 3 8 6 31 15 31 39 4 5 18 9 6 0 1 3 3 1 1 0 0 8 0 1 1 5 1 1 2 0 2 0 1 3 3 2 0 1 4 0 2 0

We will concentrate on the precincts where Whites are a majority, but still contain Hispanics, and similar communities where Blacks are a majority and still contain Hispanics. We want to do this because our **main** question is to explain if “Hispanics who live in a precinct support the black candidate to the same degree as all other ethnic groups in that precinct” - Nolan.

From observing the table, we determine to look at Precinct 11, 15, 54, 55, 82, and 84 for white majority and many Hispanics and Precinct 61, 64, 65, 81, 91, 94 and 95 for black majority and still containing many Hispanics.

We will look at the incomes of these precincts separately.





From these histograms, we can clearly see that inside these specified precincts there is a very different distribution of income. Namely, in majority black precincts the income is on average much lower than in white majority precincts.

This observation gives us reason to regress on income and race and essentially remove the info on precinct. In the next section, "Methodology and Results" this claim will be shown through model selection.

Methodology and Results

Now we will decide the best model using BIC as our model selection criteria. The model names correspond to the predictors and the interaction terms involved. For example, `bin.glm.r.i.ri` represents the binary logistic regression model where `race`, `income` and `race*income` are the predictors.

	model	BIC
1	bin.glm.p	1719.84
2	bin.glm.r	1254.78
3	bin.glm.i	1771.83
4	bin.glm.p.r	1386.18
5	bin.glm.p.i	1765.41
6	bin.glm.r.i	1297.97
7	bin.glm.p.r.i	1429.36
8	bin.glm.p.r.pr	45738.78
9	bin.glm.p.i.pi	46608.16
10	bin.glm.r.i.ri	1492.21
11	bin.glm.p.r.i.pi	35539.40
12	bin.glm.p.r.i.pr	38438.17
13	bin.glm.p.r.i.ri	1622.69
14	bin.glm.p.r.i.pi.pr	35628.89
15	bin.glm.p.r.i.pi.ri	37000.20
16	bin.glm.p.r.i.pr.ri	39014.32
17	bin.glm.p.r.i.pi.pr.ri	39014.32

We see that the two models with the two lowest BIC, are `race` as a predictor and `race` and `income` as a predictor. And we show from the ANOVA chart that either is viable.

```
## Analysis of Deviance Table
##
## Model 1: cbind(total_jackson, totalnot_jackson) ~ race
## Model 2: cbind(total_jackson, totalnot_jackson) ~ race + income
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       684     842.02
## 2       676     832.92  8   9.0987    0.334
```

Now we will interpret the `race` and `income` model:

$$\text{Jackson} = -0.56 - 1.17W - 0.73H + 2.16Bl - 1.2A - 0.4Other + .65in1 + 0.67in2 + .69in3 + .76in4 + .78in5 + .70in6 + .69in7 + .61in8$$

Looking at our model we can say that in the presence of `income` as a predictor that Blacks are positively correlated with voting for Jackson. This makes perfect sense given our EDA stacked barplot that displayed race and voting for Jackson or not.

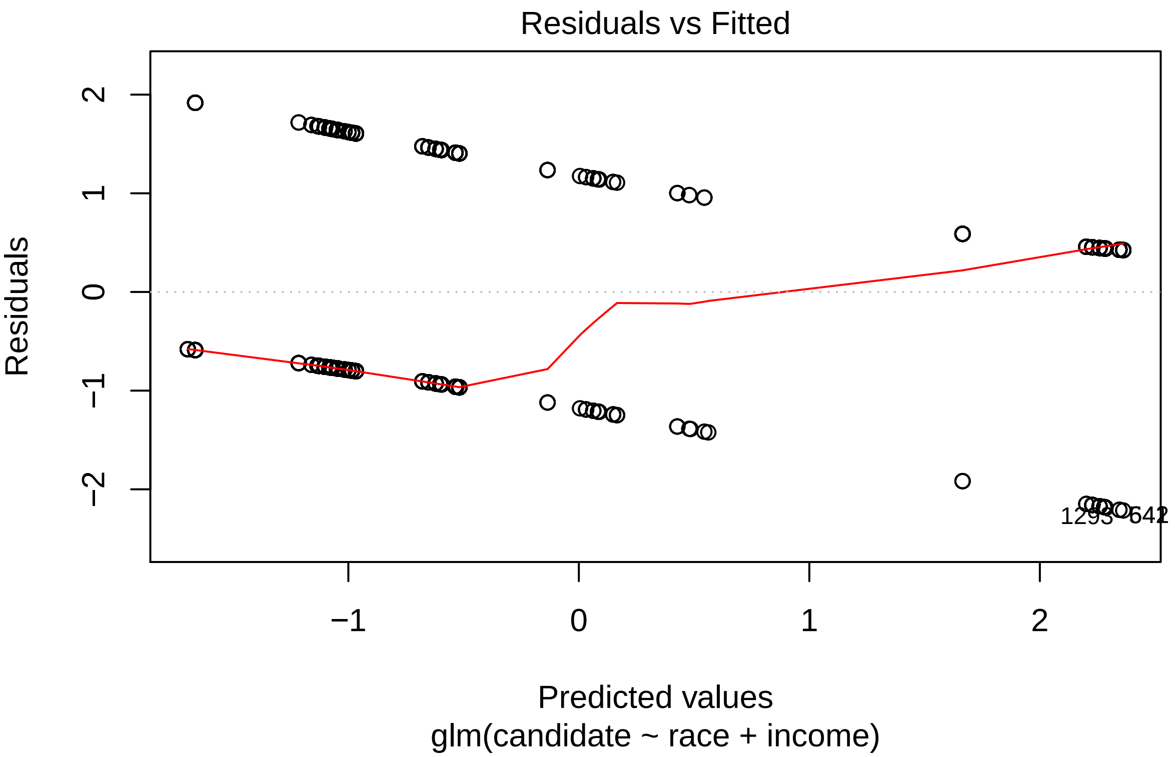
The coefficient for variable is interpreted in the presence of the rest of the variables. We must also remember that $e^{coefficient}$ is the factor by which the log odds are changed.

Let us choose the value 40 for Black and then plug this value into the fitted model. Then the relationship between voting for Jackson and the rest of the predictors for 40 as the Black value is:

$$\text{Jackson} = -0.56 - 1.17W - 0.73H + 2.16 * 40 - 1.2A - 0.4Other + .65in1 + 0.67in2 + .69in3 + .76in4 + .78in5 + .70in6 + .69in7 + .61in8$$

Discussion

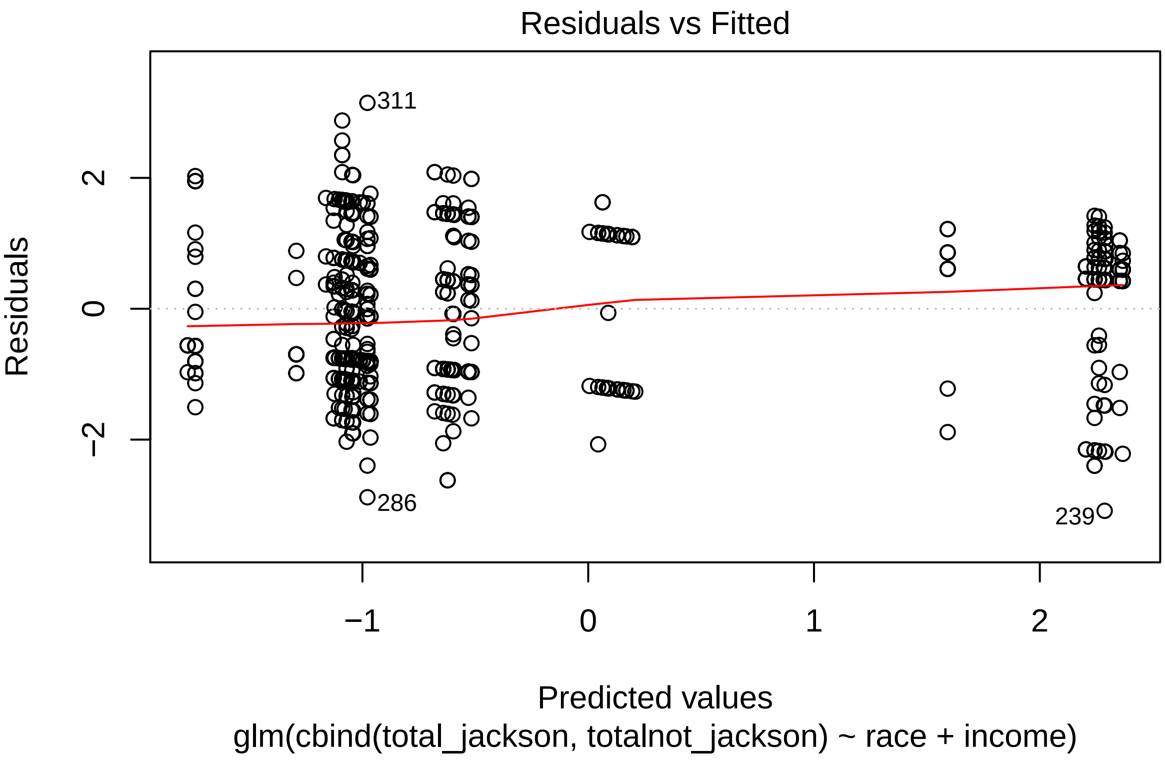
Since we are dealing with a classification problem, 0 for not Jackson and 1 for Jackson, we first try logistical regression with race and income as the two predictors. And we plot the residuals:



Here, we clearly have two curves, due to the fact that y_i is 0 or 1, so we have $-2log(\hat{\pi}_i)$ when y_i is 1 and $-2log(1-\hat{\pi}_i)$ when y_i is 0.

Let us instead try the binomial logistic regression. We must group our data by precinct, income, and race and sum for and against Jackson based on this.

First, we look at the residuals of the same model as above, but for the binomial logistic regression.



This residual fit is much better than the previous residual fit. The reason, there are so many clusters in the plots stems from race and income are both composed of levels.

Another thing to notice about the data set is the presence of Asians in the population was very sparse. The only two non-white groups are essentially Black and Hispanic. It would be interesting to study a data set where Asians were more prevalent. For this reason my model is **not** generalizable to other data sets with the same structure as this one.

Summary

We found that `race` or `race` and `income` are the best models to predict whether a vote will be for Jackson or not. This allows us to essentially drop precinct. The overall takeaway is to look across precincts rather than within precincts when determining if a vote will be for Jackson or not. Therefore Hispanics do not support the Black candidate as much as other races in their precinct. Support for the Black candidate is explained by income and race more than precinct.

References

Nolan's lecture notes:

- Lecture 05.KaiserBabies.html
- Lecture 21: Model Selection
- Lecture 25: Binomial Logistic

R packages:

- DataComputing

- tab
- dplyr
- xtable