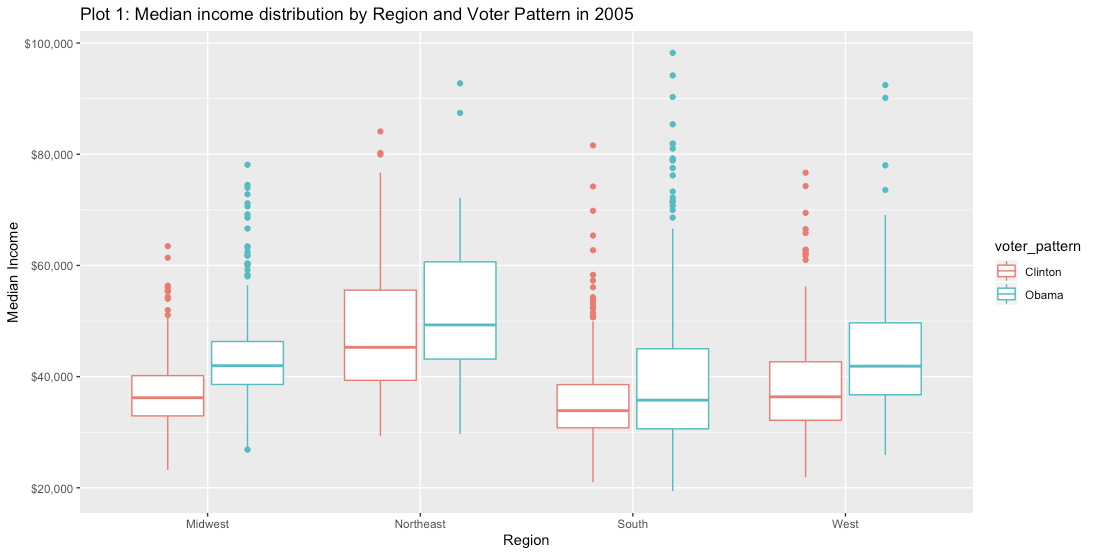
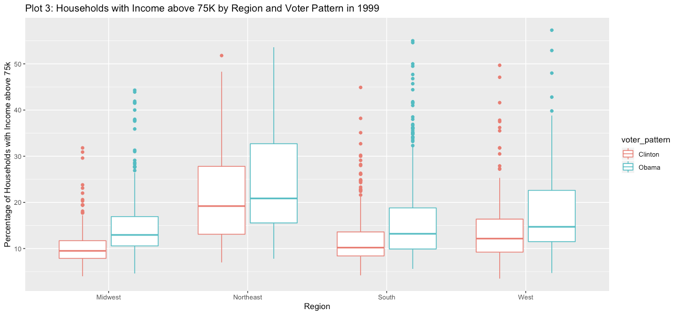
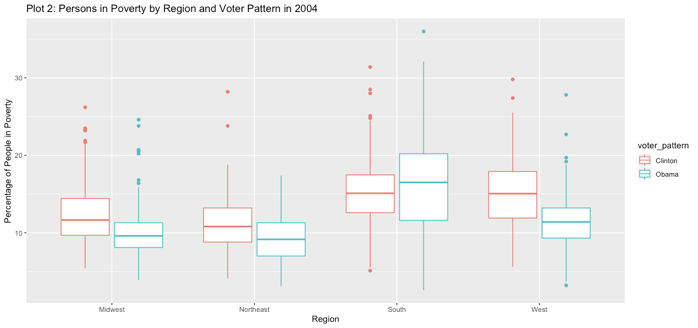
**Case : 2008 Democratic Primaries – Clinton vs. Obama**

**Data exploration**

**Formulated Question:** Does income have an influence on candidate preference, across regions?



Plot 1 indicates that counties within regions with a higher median income supported Barack Obama more than Hillary Clinton as opposed to counties within regions with a lower median income, who supported Clinton more than Obama. To analyze this trend further, we decided to analyze candidate support in high-income regions (households with income above 75K) and low-income regions (percentage of people in poverty) in 2005.

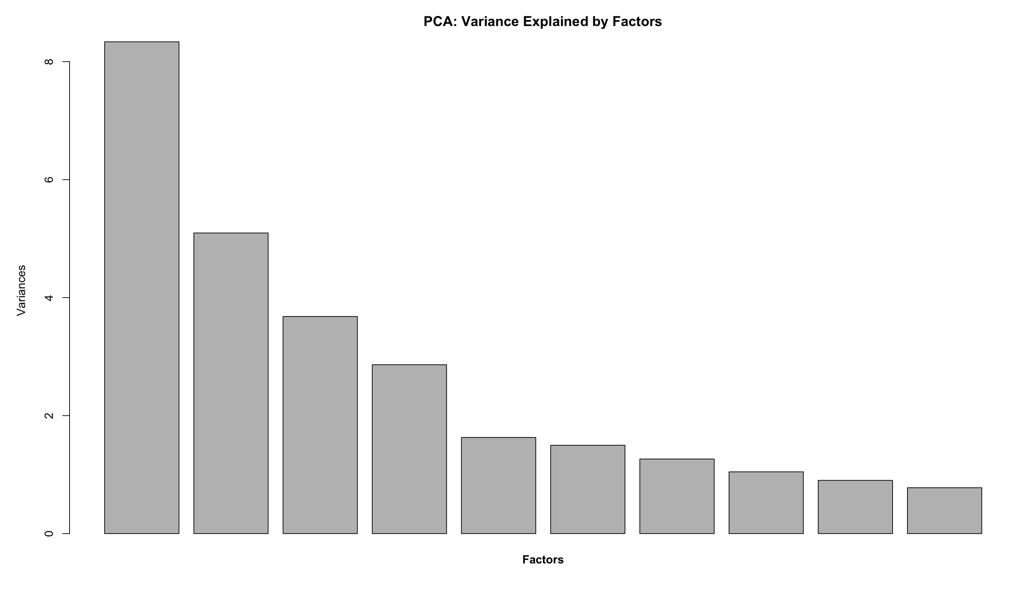


Plot 2 depicts that the percentage of persons in poverty in counties in the Midwest, Northeast and West supported Clinton more than Obama, while persons in poverty in counties in the South supported Obama more than Clinton in 2004. Plot 3 highlights that the percentage of households with income above 75K in counties in all 4 regions supported Obama more than Clinton. The three plots outline that people belonging to a lower income bracket were Clinton supporters and people belonging to higher income brackets were Obama supporters.

While more data regarding specific information of voter preference at a persons-level is required to understand this trend, we hypothesis that such a pattern could be a function of the fact that Obama was viewed as an outsider, given his subdued political presence before the primaries campaign. Hence, low income people were perhaps less trusting in him and preferred the familiar face of Clinton. The interesting insight here is depicted in Plot 2 with the people in poverty in the South supporting Obama more than Clinton. The south has a strong black community and therefore perhaps trusted Obama more, given his African American ethnicity.

However, when analyzing these 3 plots, it should be noted that no causation can be drawn because we do not have information at a voter level, i.e. we don’t have information about a voter’s preference and his/her income. We’ve drawn these conclusions from couny-level data. It is also important to consider that income information is aggregated across the entire population in a county, however, not all the people in a county voted for either Obama or Clinton. Many perhaps abstained, were perhaps ineligible or chose to vote for another party. Therefore, to draw more accurate conclusions, we’re require more specific data.

**Unsupervised learning (PCA)**



We decided to run a principle component analysis to group the data by their shared latent features. We first graphed how the variation in the data would be described given a certain number of factors; it appeared that four factors would be sufficient to accurately characterize the data points.

By looking further into the loadings of variables within each principal component, we concluded the key information that each principal component contains.

PC1 accounts for information related to desirable social characteristics such as Bachelors(education), Social Security, Income, Medicare, etc.

PC2 includes information related to undesirable social characteristics such as poverty, disability, unemployment rate, etc.

PC3 accounts for age and work status information.

PC4 includes information about ethnicity (mostly black, all other races are negatively correlated).

**Comparing models and evaluate model performance using K-fold cross validation**

Our goal is to create a model that can best predict the winning spread of Obama over Clinton using R2 and AIC values to evaluate their performance. We began by testing different regression models and comparing their performance. First, we ran a full linear regression model that included all variables and resulted in an adjusted R2 of 0.8453 and an AIC of 13903.71. Then we ran a stepwise regression, which resulted in a model with 46 variables and an adjusted R2 of 0.8308 and an AIC of 13744.06; it removed all county categorical variables as well as some of the states to rely primarily on demographic describing variables. We then set up a different model by using lasso to select variables; this resulted in a model with 31 variables, an adjusted R2 of 0.8292, and an AIC of 13746.28

In order to decide which model we should use, we compared the R2 values and the AIC of all our models. While the full model reported the highest adjusted R2, it also had a higher AIC than the other models (and in fact the difference in R2 between this model and the lasso model is quite small). In addition, the lasso model has a slightly lower adjusted R2 than the stepwise as well as a slightly lower AIC than the stepwise model, but the number of variables in lasso model is much less than that of a stepwise model; as a result, we have chosen to use the lasso model (with 1se rule) as our final model in predicting the winning margin.

After discovering that the lasso model was the most effective at predicting the winning spread, we applied a 10-fold cross validation to evaluate the model’s performance. We then used our final model to predict the winning percentages for the test sample.

We use MSE as the evaluation metrics for the 10-fold cross validation. The 10-fold cross validation MSE results are shown below:

[1,] 142.8315 [2,] 171.0646 [3,] 205.5809 [4,] 219.8708 [5,] 170.4374 [6,] 142.2448

[7,] 195.5427 [8,] 159.8097 [9,] 167.5990 [10,] 155.9426

The prediction result is shown in the R file.

**Investigate impact of changing hispanic demographic**

Hispanic:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **T value** | **Pr(>|t|)** |  |
| Hispanic\_double | -0.04670547 | 0.11343666 | -0.41173166 | 0.68058889 |  |
| Hispanic\_simple | -0.11511 | 0.06973 | -1.651 | 0.098940 | . |

Assuming the Conditional Independence Assumption (CIA) holds, the simple linear regression model shows that with a 5% increase in Hispanic demographic, the winning margin percentage for Obama over Clinton will decrease by 0.57555%. However, the “Hispanic” coefficient is not significant at 0.05 level, thus, the simple model doesn’t provide validate estimation on how increase in Hispanic population will impact Obama’s winning spread.

The improved model using double selection approach has a smaller negative coefficient than the simple model, and the coefficient is even more insignificant enough for us to make any casual inference between Hispanic demographic and Obama’s winning spread. This is because when we use double selection approach to control variables other than the treatment itself, the treatment coefficient in the double selected model reveals more accurate information on its independent impact on the dependent variable y.

In this case, neither model supports strong correlation between Hispanic demographic and Obama’s winning spread over Clinton.

**Investigate impact of changing black demographic**

Black:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **T value** | **Pr(>|t|)** |  |
| Black\_double | 0.30484316 | 0.17850569 | 1.70775036 | 0.08786853 | . |
| Black\_simple | 0.86311 | 0.04204 | 20.53 | <2e-16 | \*\*\* |

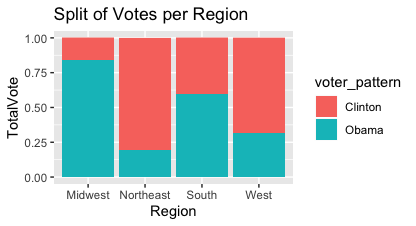
Assuming the Conditional Independence Assumption (CIA) holds, the simple linear regression model shows that with a 5% increase in Black demographic, the winning margin percentage for Obama over Clinton will increase by 4.31555%. The “Black” coefficient is significant at 0.0001 level, which indicates that the simple model provides validate estimation on how increase in Black population will impact Obama’s winning spread.

However, the improved model using double selection approach has a smaller coefficient which is not significant enough for us to make any casual inference between Black demographic and Obama’s winning spread. This is because when we use simple linear regression model, the coefficient also captured the effects from other variables to the dependent variable. When we use the double selection approach to control variables other than the treatment “Black” itself, the impacts of other variables are excluded, which leaves treatment coefficient in the double selected model to reveal more accurate information on its independent impact on the dependent variable y.

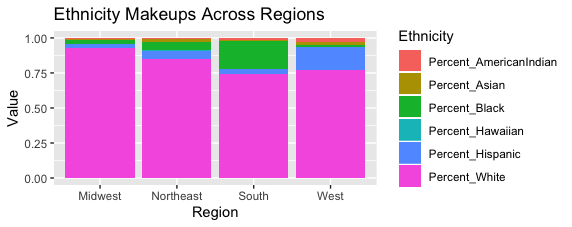
In this case, the model using double selected approach provides more accurate information. Additionally, since the coefficient is not significant enough, we are not able to make any casual inference between Black demographic and Obama’s winning margin percentage over Clinton.

**Further exploration on Obama’s voter demographics and recommendations**

We chose to analyze Obama’s voter demographics to determine where the Obama campaign should allocate their time and resources and what segments they should appeal to in order to improve their standing using graphs to illustrate. First, a simple percent bar chart shows that Obama received the smallest percentage of the total votes in the Northeast and West regions; he has a substantial influence in the Midwest which is most likely due to his popularity and good standing from his time working as senator for Illinois. Moving forward, the campaign should allocate more resources and volunteer time to the Northeast and West regions in order to help increase the percentage of votes for Obama.



Based on the chart below, we can see the majority of the Northeast and West consist of white citizens as well as black and Hispanic; given that the Hispanic population represents a large portion of the Western votes and the data in question 3 showed that there is a possible negative relationship with Hispanics voting and Obama receiving more votes, the Obama campaign should target Hispanic voters in the West (as well as the Northeast). This will hopefully increase Obama’s performance in these regions.



The insights discovered in question 1 can also be used to influence this decision; in our initial data visualization we found that Obama support is lacking in the more impoverished counties and thus focusing on campaigning in the Northeast and West targeting people of lower incomes would be beneficial to the campaign.