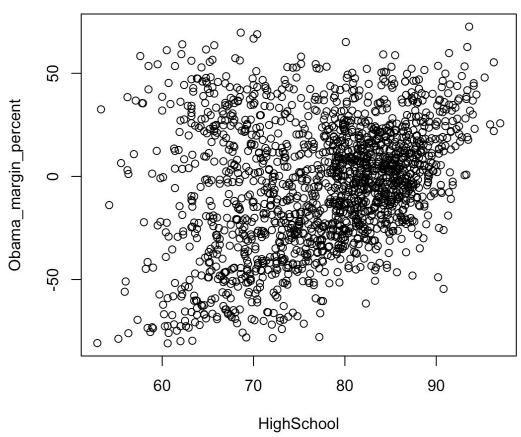
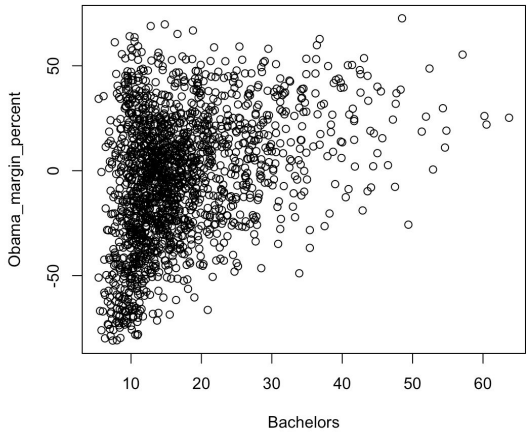
Data Science - 2008 Democratic Primaries - Clinton vs. Obama

Overview

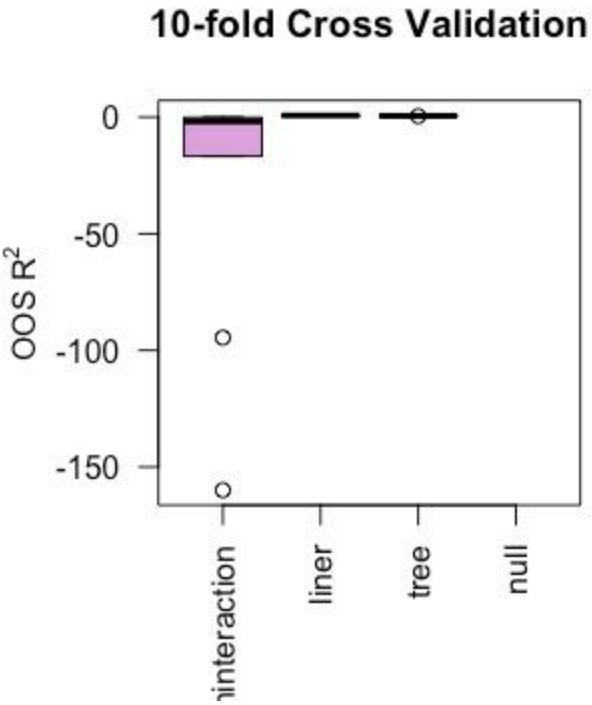




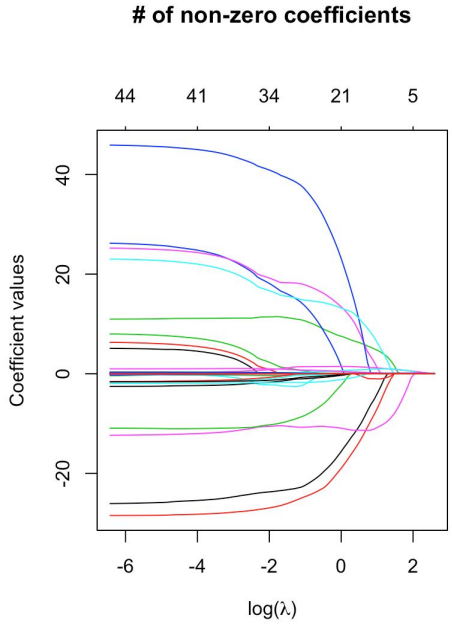
Our group picked two variables that we are interested in - the county's population's education level and Obama margin percent. We plotted obama margin percent against bachelors and high school, and R returned the above two graphs. From these two pictures, we concluded that holding everything else constant, for the counties with a more educated population, the winning spread of Obama over Clinton are expected to be larger. We got this conclusion because the bottom right of both graphs are empty, which means that most counties with a high proportion of educated people did not vote for Clinton. The difference in the location of dot mass is because the proportion of people with a high school degree is larger than the proportion of people with a bachelor degree. With this information, Clinton and Obama can decide which population group to target for the election campaign.

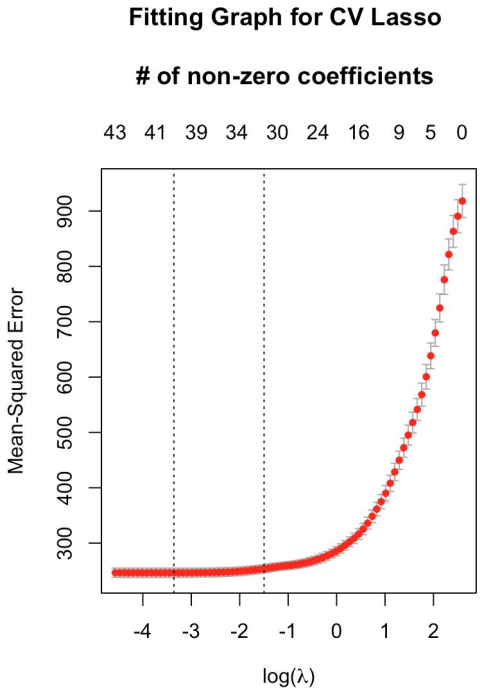
Core task: We want to select a set of meaningful variables and find the best fitted model based on these variables. And we will use two steps to find our final choice, including model choice and variables choice.

Model Description: First, we use four models, including linear interaction model, linear model, tree classification model and null model. We began with K-fold cross validation and cross validate linear regression model, linear interaction model, null model and tree classification model within these 10 folders. In order to choose the best fitted model, we calculate R-square of 10 folders to compare the results among four models. We got the Out of sample R-square of linear regression model is 0.804, the OOS R-square of linear interaction model is -0.120, the OOS R-square of null model is 0, the OOS R-square of tree classification is -0.123. Based on all these R-square results, linear regression model has the highest R-square which means this is the best model. Then we will use lasso to choose variables.

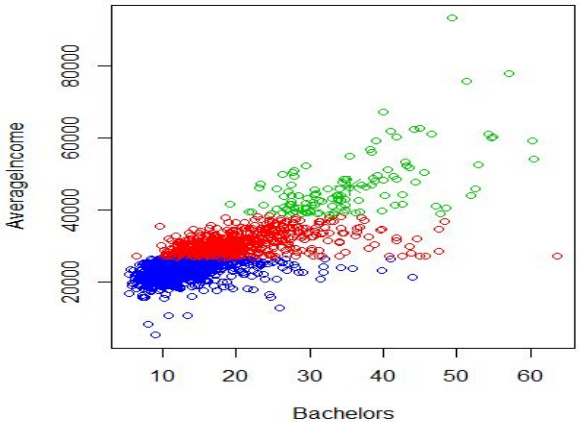


Model Validation: Based on the R-square results, we found linear regression model had the highest R-square and is the best model among all four options, so we start with linear regression model. To evaluate linear regression model, we use Lasso to select the best set of variables. After running Lasso, we found RetiredWorkers, Medicare, Pop and Age35to65 are not as effective as the other variables, so we delete these variables and get the final model.





Prediction: We use the model from above to run prediction in the test data set and get the prediction for the Obama_margin_percent on each observation.



We used the k-mean to exploring the data among bachelors and average income. The x-axis is the percentage of bachelor or higher and y-axis stands for the average income. From we can see from the plot above, there are 3 clusters with blue, green and red.

The blue cluster represents a group of people with a lower level of education and a lower average income. The red cluster represents people that have both an average income and average education. We guess these 3 clusters may stand for the different communities. The blue one may be the “bad” community where a group of people with a lower level of education and a lower average income, and the red one is average where gathering people that have both an average income and average education. Finally, the green cluster represents the good community where people with higher than average income and education.

We run a simple regression only between Obama_margin_percent and Hispanic, and we find the p-value of Hispanic is not significant, so we just contain all the variables except those defined by Lasso. Finally, we can conclude that when if the percentage of hispanic increase by 5%, the winning spread for Obama over Clinton will increase 0.2467*5 = 1.2335.

```
Call:
glm(formula = Obama_margin_percent ~ Hispanic + MalesPer100Females +
    AgeBelow35 + White + Black + Asian + AmericanIndian + Hawaiian +
    HighSchool + Bachelors + Poverty + IncomeAbove75K + MedianIncome +
    AverageIncome + UnemployRate + ManfEmploy + SpeakingNonEnglish +
    MedicareRate + SocialSecurity + SocialSecurityRate + Disabilities +
    DisabilitiesRate + Homeowner + SameHouse1995and2000 + PopDensity +
    LandArea + FarmArea, data = election_data_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-80.150  -12.319    0.511   13.095   77.540

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -9.284e+01  3.601e+01  -2.578 0.010023 *
Hispanic             2.467e-01  1.291e-01   1.911 0.056124 .
MalesPer100Females   1.268e-01  5.666e-02   2.238 0.025361 *
AgeBelow35           4.701e-01  1.790e-01   2.626 0.008723 **
White               -9.651e-01  2.831e-01  -3.409 0.000666 ***
Black                7.249e-01  2.777e-01   2.610 0.009123 **
Asian               -1.220e+00  4.565e-01  -2.672 0.007608 **
AmericanIndian      -1.072e+00  3.388e-01  -3.164 0.001584 **
Hawaiian            -2.513e+00  4.510e+00  -0.557 0.577446
HighSchool           1.378e+00  1.237e-01  11.143  < 2e-16 ***
```

```
Bachelors             1.105e+00  1.439e-01   7.682 2.63e-14 ***
Poverty              -1.796e+00  3.208e-01  -5.598 2.52e-08 ***
IncomeAbove75K       -1.020e+00  2.623e-01  -3.887 0.000106 ***
MedianIncome         -3.227e-04  2.112e-04  -1.528 0.126641
AverageIncome         4.038e-04  1.466e-04   2.755 0.005939 **
UnemployRate          6.928e-01  3.722e-01   1.861 0.062884 .
ManfEmploy           -1.201e-01  7.557e-02  -1.589 0.112144
SpeakingNonEnglish    1.620e-01  1.598e-01   1.014 0.310888
MedicareRate          7.595e-04  1.725e-04   4.404 1.13e-05 ***
SocialSecurity       -1.972e-04  2.468e-05  -7.990 2.45e-15 ***
SocialSecurityRate   -1.414e-03  2.931e-04  -4.823 1.54e-06 ***
Disabilities          5.117e-04  8.600e-05   5.950 3.25e-09 ***
DisabilitiesRate     -2.231e-03  6.877e-04  -3.244 0.001200 **
Homeowner             4.897e-01  1.088e-01   4.500 7.26e-06 ***
SameHouse1995and2000  4.857e-01  1.089e-01   4.460 8.74e-06 ***
PopDensity           -8.839e-04  2.608e-04  -3.389 0.000716 ***
LandArea              2.354e-03  3.884e-04   6.061 1.66e-09 ***
FarmArea             -1.460e-03  1.881e-03  -0.776 0.437700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 376.004)

    Null deviance: 1595745  on 1736  degrees of freedom
Residual deviance:  642591  on 1709  degrees of freedom
AIC: 15259

Number of Fisher Scoring iterations: 2
```

We first confirm the variables should be deleted from the outcome of lasso, which are mentioned in Q2. Then we use forward step-wise analysis to control the omitted variables step by step, and keep the controls variables whose coefficient are significant different from 0. Finally we find the coefficient of black in our model is 1.648. Therefore we can conclude that when if the percentage of black increase by 5%, the winning spread for Obama over Clinton will increase 1.648*5 = 8.24.

```
Call:
glm(formula = Obama_margin_percent ~ Black + HighSchool + Poverty +
    SpeakingNonEnglish + SocialSecurity + Disabilities + PopDensity +
```

```
   LandArea, data = election_data_train)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-84.694 -12.927   0.058  13.580  95.866

Coefficients:
                    Estimate Std.      Error t value           Pr(>|t|)
(Intercept)         -1.534e+02     8.526e+00 -17.995     < 2e-16 ***
Black                1.648e+00     4.021e-02  40.985     < 2e-16 ***
HighSchool           1.898e+00     8.905e-02  21.318     < 2e-16 ***
Poverty             -1.376e+00     1.701e-01  -8.089     1.12e-15 ***
SpeakingNonEnglish   4.651e-01     6.236e-02   7.458     1.38e-13 ***
SocialSecurity      -2.342e-04     2.294e-05 -10.212     < 2e-16 ***
Disabilities         5.485e-04     8.394e-05   6.535     8.37e-11 ***
PopDensity          -1.106e-03     2.455e-04  -4.504     7.11e-06 ***
LandArea             2.008e-03     3.630e-04   5.532     3.64e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 414.2864)

   Null deviance: 1595745  on 1736  degrees of freedom
Residual deviance:  715887  on 1728  degrees of freedom
AIC: 15408

Number of Fisher Scoring iterations: 2
```

Insights: Our team will focus on Obama. One insight we gained from our analysis is that areas with higher education seem to favor Obama. This means that people with a lower education will not tend to vote for Obama. Therefore, Obama’s campaign team must focus efforts and resources on reaching out to these people and convincing them to vote for Obama.
Another insight we gained from our analysis is that black voters seem more likely to vote for Obama than hispanic voters. The campaign team must therefore also focus its attention to the hispanic community because they constitute the fastest growing population in the US