Modeling HW Part 1

Kerui Cao

9/26/2019

Data analysis

1992 presidential election

The folder `nes` contains the survey data of presidential preference and income for the 1992 election analyzed in Section 5.1, along with other variables including sex, ethnicity, education, party identification, and political ideology.

1. Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.

Before dive into analysis, we have to reorganize and clean the data, for example we can notice that there are a lot missing values,we have to delete them or impute them.

Table 1: Most Missing Value Variables

	icpsr_cty	regis	perfin2	real_ideo	parent_party
Number of Unique Value	1	1	1	8.000	6.000
Number of Missing Value	1222	1222	1222	243.000	243.000
% of Missing Value	1	1	1	0.199	0.199

Table 2: Least Missing Value Variables

	presadm	age_10	age_sq_10	white	vote_rep
Number of Unique Value	1	73	73	2	2
Number of Missing Value	0	0	0	0	0
% of Missing Value	0	0	0	0	0

Table 3: Least Unique Value Variables

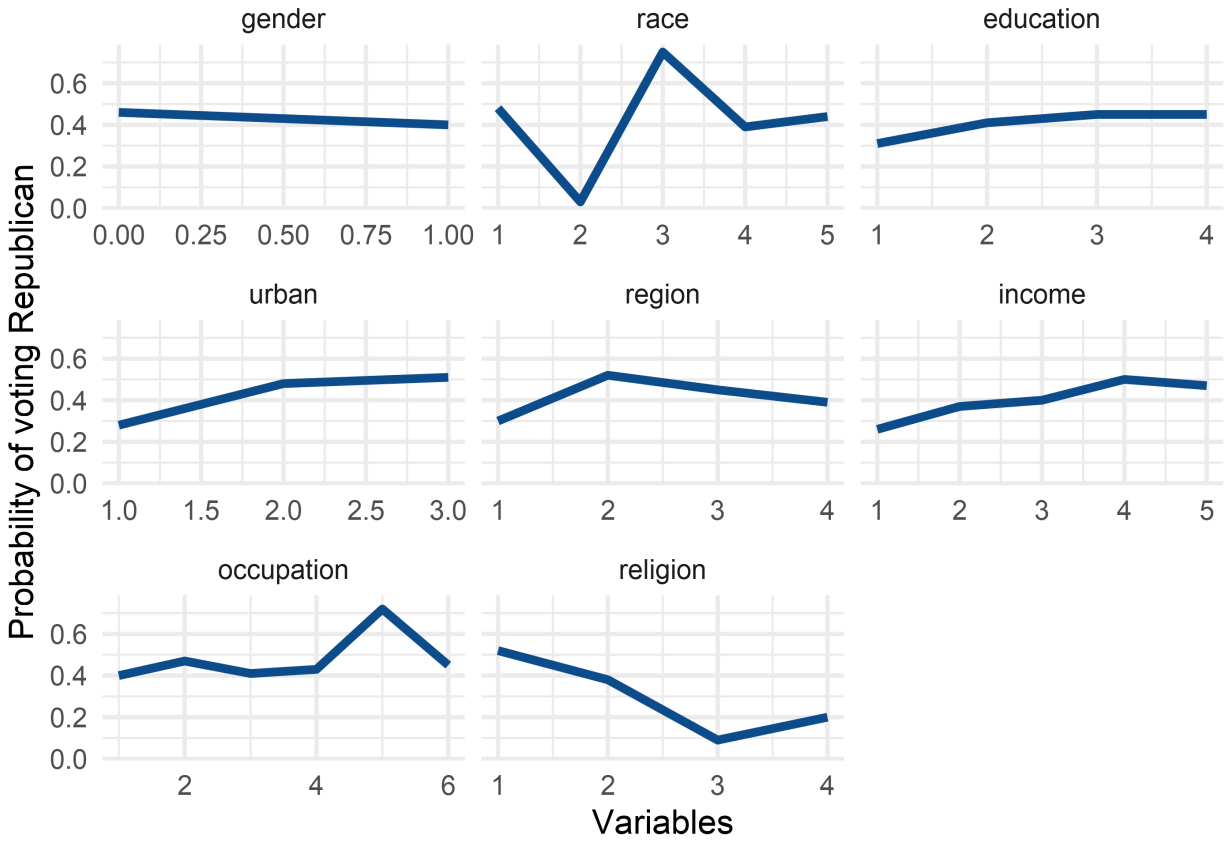
	icpsr_cty	regis	perfin2	year	vote
Number of Unique Value	1	1	1	1	1
Number of Missing Value	1222	1222	1222	0	0
% of Missing Value	1	1	1	0	0

According to the table above, we can see that for some variables, they are all missing values, and for some variables, they have only one value, for now we will delete all of these kind of variables, as for variables that is partly missing, we will deal with them later after predictors being decided.

Take a look at the data, there are many categoriacal variables, so we have to transform them into numeric data.

As for the dependent variablel, there is no variable in data that shows the final vote for each respondents, so we can only use the result for prevote as the estimate of final vote result. and according to the formation of data set nes5200_dt_s, the value of presvote can only be “democrat” or “republican”, 1 represents democrat, 2 represents republicans, in order to construct logistic regression, we substract prevotes by 1, so here 0 represents democrat, 1 represents republicans.

After the precess above, we can begin to analyze the data, first we plot some intersted variables against the dependant variable:

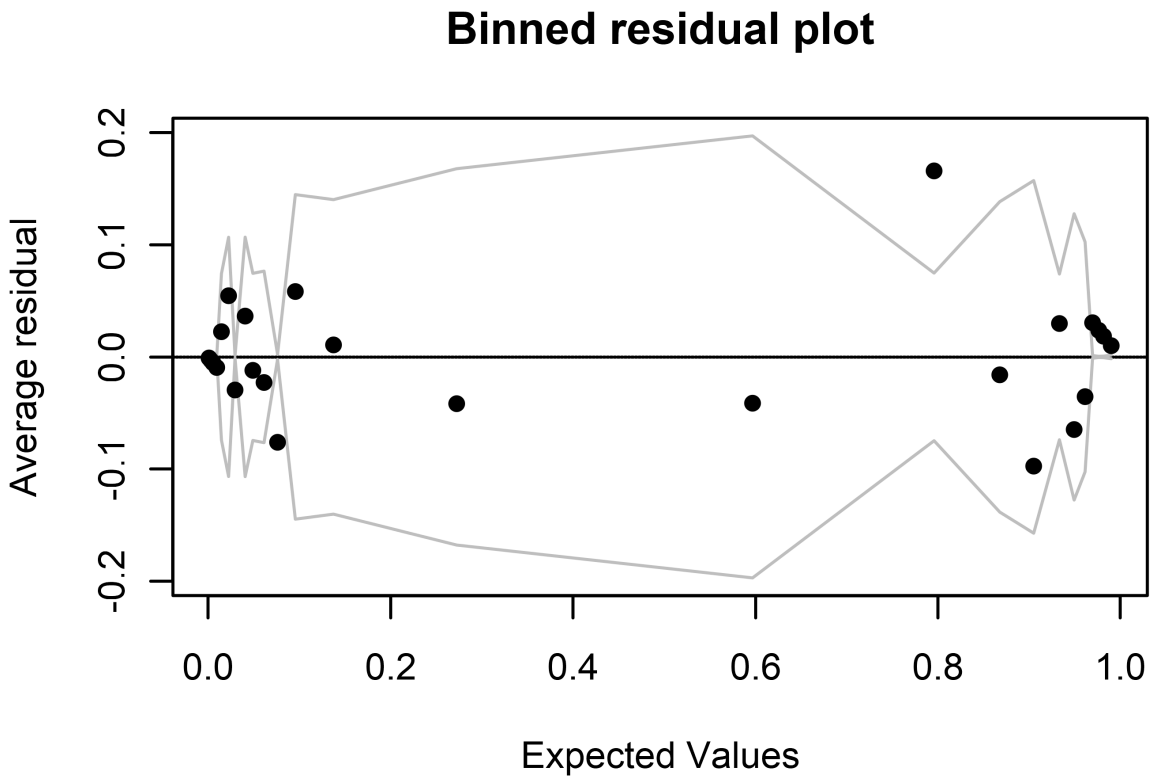


We can see that the probability of voting Republican varies a lot for each variables we choose. Now we can construct logistic regression, as for transformation of variables, because most of the input variables are categorical variables, so it is not helpful to transform categorical variables, so does interaction between categorical.

```
##
## Call:
## glm(formula = pre_vote ~ age + gender + factor(race) + factor(r_education) +
##     factor(urban) + factor(region) + factor(income_level) + factor(r_occup) +
##     factor(religion) + factor(martial) + factor(par_id_3) + factor(father_party) +
##     factor(mother_party), family = binomial(link = "logit"),
##     data = da)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.52546  -0.30664  -0.08796   0.28236   2.92928
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -4.6773103  1.6483345  -2.838  0.00455 **
## age                    0.0134399  0.0129279   1.040  0.29852
## gender                -0.0103552  0.3750280  -0.028  0.97797
## factor(race)2         -2.8953728  0.9647313  -3.001  0.00269 **
```

```
## factor(race)3          2.0380346  1.8477298   1.103  0.27003
## factor(race)4          0.0004864  1.0144746   0.000  0.99962
## factor(race)5          1.6884572  0.8505319   1.985  0.04712 *
## factor(r_education)2  -0.1114266  0.9010321  -0.124  0.90158
## factor(r_education)3  -0.2891096  0.9521841  -0.304  0.76141
## factor(r_education)4   0.5354012  0.9665532   0.554  0.57963
## factor(urban)2        -0.0188918  0.4458202  -0.042  0.96620
## factor(urban)3         0.1029981  0.4879832   0.211  0.83283
## factor(region)2        0.5085193  0.5030997   1.011  0.31213
## factor(region)3        0.7829948  0.5200502   1.506  0.13217
## factor(region)4       -0.0765439  0.5560230  -0.138  0.89051
## factor(income_level)2  0.7490035  0.8130482   0.921  0.35693
## factor(income_level)3  0.4681915  0.7638749   0.613  0.53993
## factor(income_level)4  0.2654111  0.7999655   0.332  0.74006
## factor(income_level)5 -1.2470272  0.9512761  -1.311  0.18989
## factor(r_occup)2       0.7872371  0.4766595   1.652  0.09862 .
## factor(r_occup)3       0.5133552  0.4850627   1.058  0.28991
## factor(r_occup)4      -1.8599863  1.3100138  -1.420  0.15566
## factor(r_occup)5       0.8344592  1.0728789   0.778  0.43670
## factor(r_occup)6       0.0152706  0.6317855   0.024  0.98072
## factor(religion)2      0.3155637  0.4322414   0.730  0.46535
## factor(religion)3     -2.0319529  1.2692955  -1.601  0.10941
## factor(religion)4     -1.1791072  0.5609577  -2.102  0.03556 *
## factor(martial)2      -0.7392238  0.5860365  -1.261  0.20717
## factor(martial)3      -0.0134564  0.5356949  -0.025  0.97996
## factor(martial)4       2.3407735  1.0271093   2.279  0.02267 *
## factor(martial)5      -0.0961872  0.7274278  -0.132  0.89480
## factor(martial)7       0.6938349  1.2234991   0.567  0.57065
## factor(par_id_3)2      2.5144137  0.5843057   4.303 1.68e-05 ***
## factor(par_id_3)3      5.6285996  0.4557752  12.350  < 2e-16 ***
## factor(father_party)2 -1.0775787  0.6746624  -1.597  0.11022
## factor(father_party)3  0.2198332  0.6137052   0.358  0.72019
## factor(mother_party)2  0.3720529  0.6322568   0.588  0.55623
## factor(mother_party)3  0.4877870  0.6268564   0.778  0.43648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 904.90  on 661  degrees of freedom
## Residual deviance: 288.78  on 624  degrees of freedom
## AIC: 364.78
##
## Number of Fisher Scoring iterations: 7
```

2. Evaluate and compare the different models you have fit. Consider coefficient estimates and standard errors, residual plots, and deviances.



3. For your chosen model, discuss and compare the importance of each input variable in the prediction.

For the result of regression, we can see that race and party identity are important to their voting behavior.