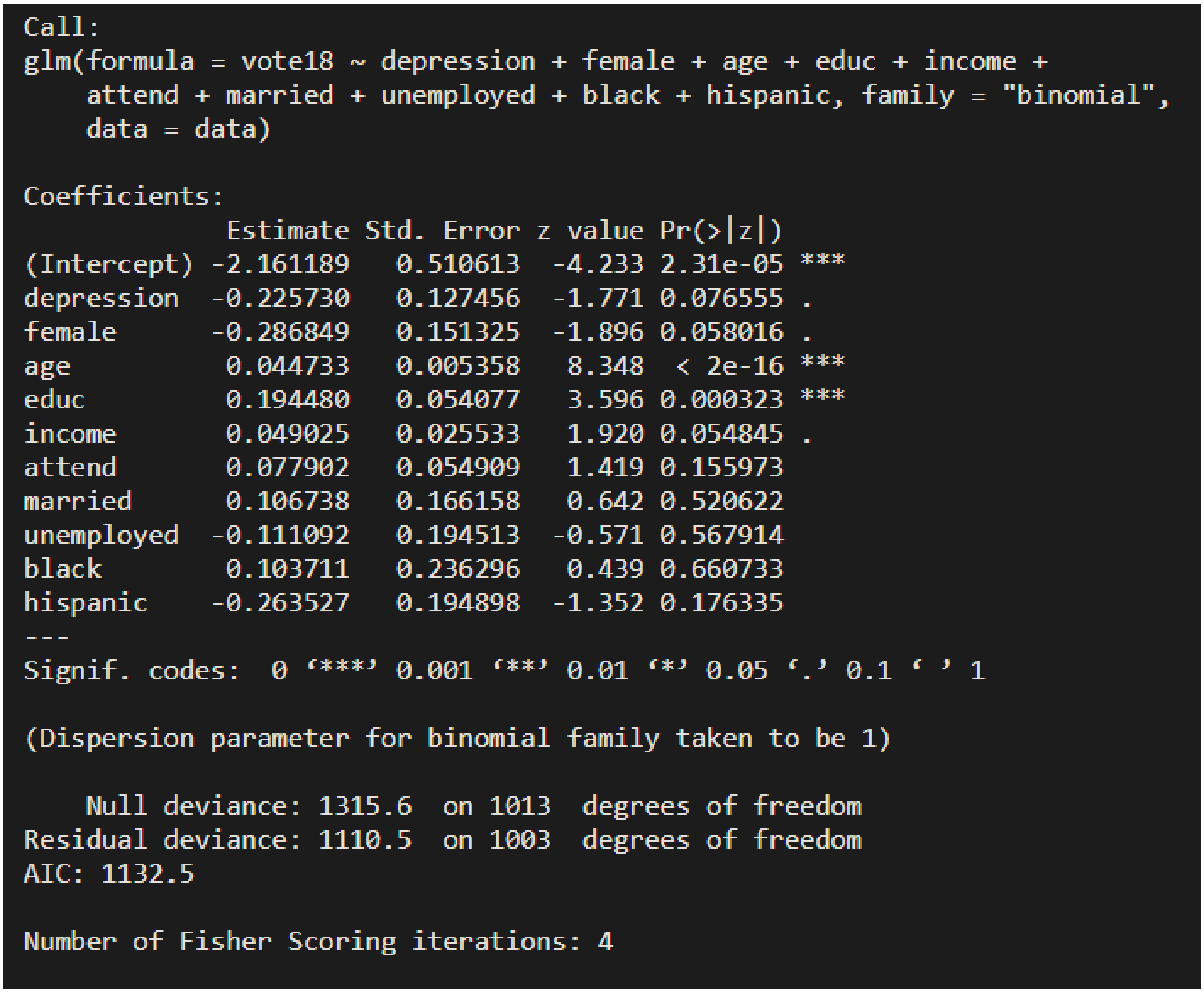
Nick Hemauer
PLSC 597
11/9/2023

Assignment 3

1. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZABHCA

The R file included has the results replicated. Below is the table that matches the table in the paper. Voter turnout is the target variable of interest.



```
Call:
glm(formula = vote18 ~ depression + female + age + educ + income + 
    attend + married + unemployed + black + hispanic, family = "binomial", 
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.161189   0.510613  -4.233 2.31e-05 ***
depression  -0.225730   0.127456  -1.771 0.076555 .  
female      -0.286849   0.151325  -1.896 0.058016 .  
age          0.044733   0.005358   8.348  < 2e-16 ***
educ         0.194480   0.054077   3.596 0.000323 ***
income       0.049025   0.025533   1.920 0.054845 .  
attend       0.077902   0.054909   1.419 0.155973    
married      0.106738   0.166158   0.642 0.520622    
unemployed  -0.111092   0.194513  -0.571 0.567914    
black        0.103711   0.236296   0.439 0.660733    
hispanic    -0.263527   0.194898  -1.352 0.176335    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1315.6  on 1013  degrees of freedom
Residual deviance: 1110.5  on 1003  degrees of freedom
AIC: 1132.5

Number of Fisher Scoring iterations: 4
```

2. I have used K-Nearest Neighbors clustering. In doing so, I split my X values into covariates I was interested in clustering on. I clustered on sex, age, education, income, black, and hispanic variables. I did this because they likely have a higher level of multicollinearity, and they all have been shown to be significantly predictive of voter turnout. Prior to clustering I scaled my data and created a Scree plot which identified 5 clusters to be an ideal clustering count.

3. Following this I clustered and compared the R2 and MSE of both the training and test logistic regressions (dependent variable = voter turnout). I found the MSE went from .29 to .28. And the R2 went from -.27 to -.23. I also found that the predictive accuracy went from 71% to 68%. Therefore, I can conclude that clustering is likely not worth the hassle in the case of this example.