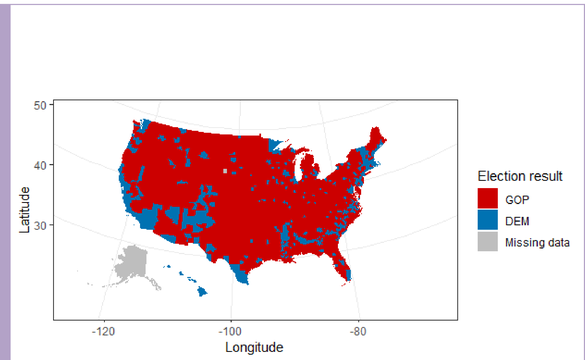
STA223 Project 1: Analysis of 2016 United States presidential election's result

Yunan Hou
Master's Student in Biostatistics
University of California at Davis

Introduction

United States presidential election's result is always very impactful and worthy to analyze. Here we want to know which demography and economic factors are significantly crucial to the 2016 election result and describe the common characteristic of the counties which won by the Democratic Party(DEM), or the Republican Party(GOP). Since our response variable, the election result is binary type (1 is DEM's win, 0 is GOP's win), we choose logistic regression to build a general linear regression model to achieve our goal.

The dataset is collected by Jia et al. [1]. It includes 3111 counties in the 48 contiguous United States (except Oglala Lakota County, South Dakota) and Hawaii and 24 variables of the election result, demography and economic factors in 2012 and 2016 (etc. unemployment rate, median income and population). Appendix Table 3 shows the bachelor rate, median income, population, civilian labor force, and the unemployment rate all have a high correlation with themselves in the different years. After the colinear check, we delete all the 2012's variables except for 2012's election result and use the rest of the predictors to build the first model.



Method

- Model Building

Since our y is a binary response variable, we need to use logistic regression with the binomial family to build a general linear regression model to achieve our goal.
The model can be formally written as

$$P(Y=1|X=x_i)=\frac{1}{1+e^{-\beta^\intercal x_i}}, \quad \hat{y}_i=\begin{cases}1, & \text{if } P(Y=1|X=x_i)\geq\frac{1}{2}\\ 0, & \text{otherwise}\end{cases}$$

- Model Selection

BIC attempt to resolve the model selection problem by introducing a penalty term for the number of parameters in the model, which formally is

$$BIC(m)=D(m)+m\log n$$

- Goodness-of-fit Analysis

Pearson and deviance criteria

$$P=\sum_{i=1}^{n}r_{iP}^2=\sum_{i=1}^{n}\frac{y_i-\hat{y}_i}{\widehat{Var}(y_i)}, \quad D^*(y,\hat{\mu})=2\{l(y,y)-l(\hat{\mu},y)\}$$

- Model Diagnostics

Leverage points and Cook's Distance

$$W=\text{diag}\{g'(\mu_i)^2V(\mu_i)\phi\}^{-1}, \quad H=W^{\frac{1}{2}}X(X^\intercal WX)^{-1}X^\intercal W^{\frac{1}{2}}=(h_{ij})_{1\leq i,j\leq n}.$$

$$D_i=\frac{e_i^2}{p\hat{\sigma}^2}\frac{h_{ii}}{(1-h_{ii})^2},$$

Results

.

Coefficients	Estimate	Standard Error	z value	P- value
Intercept	3.073e+01	4.589e+00	6.696	2.14e-11 ***
dem_2012	3.706e-04	2.523e-05	14.690	< 2e-16 ***
gop_2012	-6.404e-04	4.412e-05	-14.516	< 2e-16 ***
BachelorRate2016	1.600e+01	1.383e+00	11.564	< 2e-16 ***
MedianIncome2016	-1.653e-03	1.675e-04	-9.866	< 2e-16 ***
POP_ESTIMATE_2016	-9.515e-05	1.298e-05	-7.330	2.30e-13 ***
Civilian_labor_force_2016	3.002e-04	3.111e-05	9.651	< 2e-16 ***
Employed_2016	-3.506e+01	4.986e+00	-7.032	2.04e-12 ***
Urban_influence_code_2013	-1.393e-01	2.948e-02	-4.724	2.31e-06 ***

Table 1: Fit Result of Reduced Model

Index	county fips	Whether reversed	$\lvert r_{iP}\rvert > 9$	Leverage points	Cook's distance
136	44003	True	True	False	0.0332
441	4027	True	True	False	0.0496
784	39093	True	True	False	0.0327
795	36103	True	True	False	0.3391
1174	13067	True	True	False	0.1300
2146	49035	True	True	True	0.1879
2299	12103	True	False	True	0.1182
2754	6059	True	False	True	2.8366
3063	42049	True	True	False	0.0340

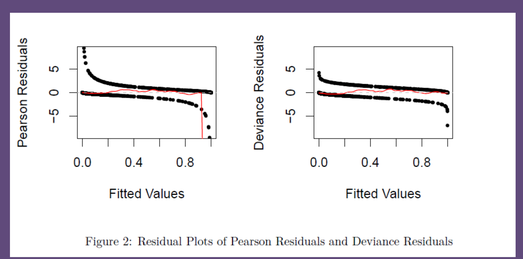


Figure 2: Residual Plots of Pearson Residuals and Deviance Residuals

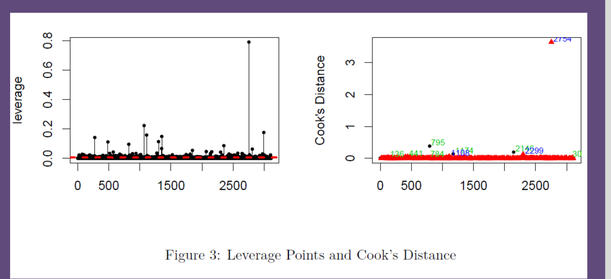


Figure 3: Leverage Points and Cook's Distance

Conclusion

As a result, the fitness of the final model is not too bad. However, the leverage test and Cook Distance found dozens of data points with great impact, of which 3 points significantly affected the model in two tests at the same time; at the same time, the fit residual of several points is very large, indicating that they are difficult to be explained by the model. The union of the two sets of points is 9 counties listed in the table below, it is interesting that all of them turn over to another party in 2016, from whom they chose in 2012.

However, after closely checking these data points, we cannot consider them to be outliers or points of numerical errors and delete them. Instead, this may be because our data set failed to include some important independent variables that can affect the results of elections, and could not accurately describe why some counties changed their supporting parties in 2016.

Therefore, it is difficult to say that our final model can be used to predict election results. In the future, we hope to include more potential influence factors.