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

Yunan Hou

Introduction and Data summary

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.

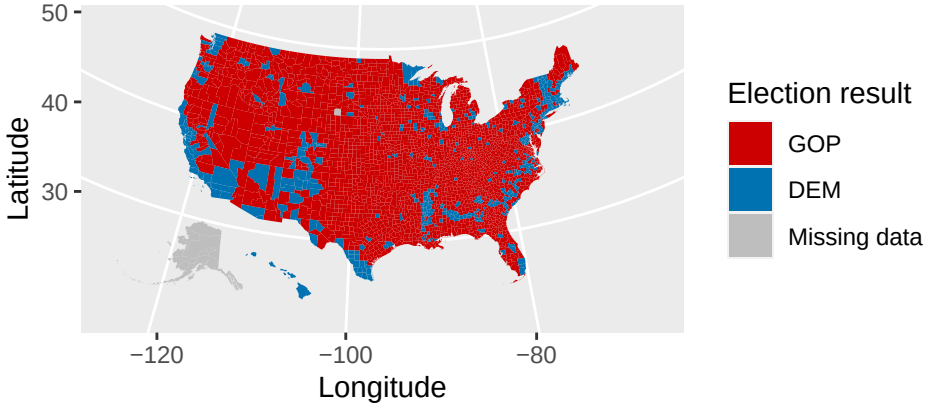

Figure 1: 2016 Presidential Election Results by County. Red represents the Republican Party(GOP) won, blue means Democratic Party(DEM) won, and white is the counties we lack data.

Model Building and Selection

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

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

where $x \in \mathcal{R}^m$ is the predictor and $\beta \in \mathcal{R}^m$ is the regression coefficient. For each state, we consider it as a Bernoulli experiment (i.e., $n_i = 1, \forall i$), in which case we have the ungrouped data.

From Table 1 in Appendix, we can see that not all predictor variables are significant. Figure 1 in the Appendix shows how the fit result looks like. To find out which factors are significantly important, we use BIC to do the model selection, since our goal is to estimate the regression coefficient rather than predicting. BIC attempt to resolve the model selection problem by introducing a penalty term for the number of parameters in the model, which formally is $BIC(m) = D(m) + m \log n$,where $D(m)$ is the deviance of a model that includes m predictors.

Then, for each state, our reduced model includes the predictors as follow: Democrats' votes in 2012, Republican Party votes in 2012, bachelor rate in 2016, median income in 2016, the population in 2016, the proportion of civilian labor force in 2016, the employment rate in 2016 and urban influence code in 2013.

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

Goodness-of-fit Analysis

Then, to do goodness-of-fit analysis on the reduced model, we check whether the residual have any kind of pattern. It is the necessary way to find whether higher-order predictors are needed. We use both Pearson and deviance criteria, which are defined as

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

.

The obtained Pearson and deviance residuals plot and boxplots from the logistic regression model are shown in the following figure 2 and Appendix Figure 2 respectively, together with the Nadaraya-Watson Kernel estimators fitted to the residuals. Notice that, for easier observation, we limit the y-axis for -9 to 9. However, there is a strong fluctuation occurs on the tail of the Pearson residual plot, which is caused by 7 extreme value points with |Pearson residual| > 9. Table 3 lists these points and we will discuss them later. From the residual plots, summary statistics, we can see that the main body of the Pearson and deviance residuals plot visually looks similar to each other. Besides, most of the residuals against fitted values show no pattern based on the smoothed line, indicating a sign of good-of-fit.

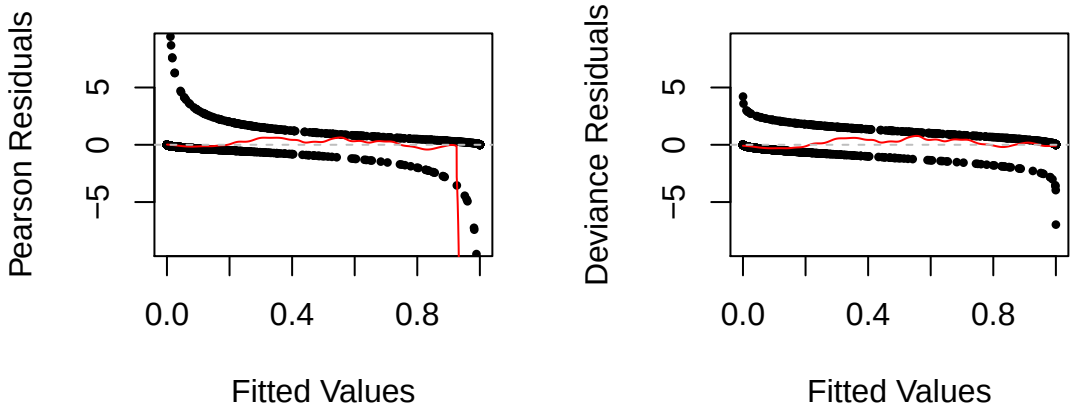

Figure 2: Residual Plots of Pearson Residuals and Deviance Residuals

For more quantitative detection, we use runs test to check whether there are significant patterns in the residuals plots(see in Appendix Figure 3 and Table 4). For both Pearson and deviance residuals, their p-values are 0.4197. Thus, we cannot reject the null-Hypothesis that there are no systematic

patterns in the residuals. Meanwhile, from the run test's plot, we cannot find systematic patterns. Thus, there is no evidence shows our model is lack-of-fit.

Model Diagnostics

We notice that there are points in the Pearson residual which have extreme value. For the next step, we check whether outliers and influential points exist by using Leverage points and Cook's Distance.For the Leverage points, we calculate

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

Then, the i_{th} observation is suspected if $h_{ii} > 2Trace(H)/n$.

Then, for the i_{th} subject, the Cook's Distance is a measure of the influence of it on the regression relation,

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

where $e_i = y_i - \hat{y}_i$ is the (standard) residual and $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2$.

In the Leverage plot, the red line is the threshold $\frac{2m}{n}$. We find 329 points that are above this threshold. In the Cook's Distance plot, the red triangles indicate the Leverage point, where the top-3 is shown by their index in blue. Meanwhile, we use green text to show the 7 points who have extreme value of Pearson residuals (one of them is also the top-3 points above). We will discuss these typically suspicious points in the next section.

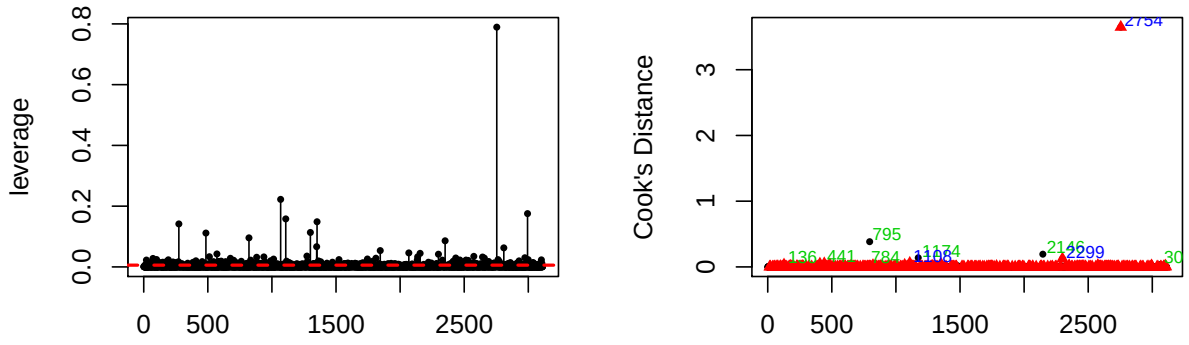

Figure 3: Leverage Points and Cook's Distance

Results and Discussion

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.

Index	county fips	Whether reversed	$\|r_{iP}\| > 9$	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

Table 3: Summary of Potential Points of Influence or Outliers

Reference

[1] Jia, Junteng, and Austion R. Benson. "Residual correlation in graph neural network regression." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.

Appendix

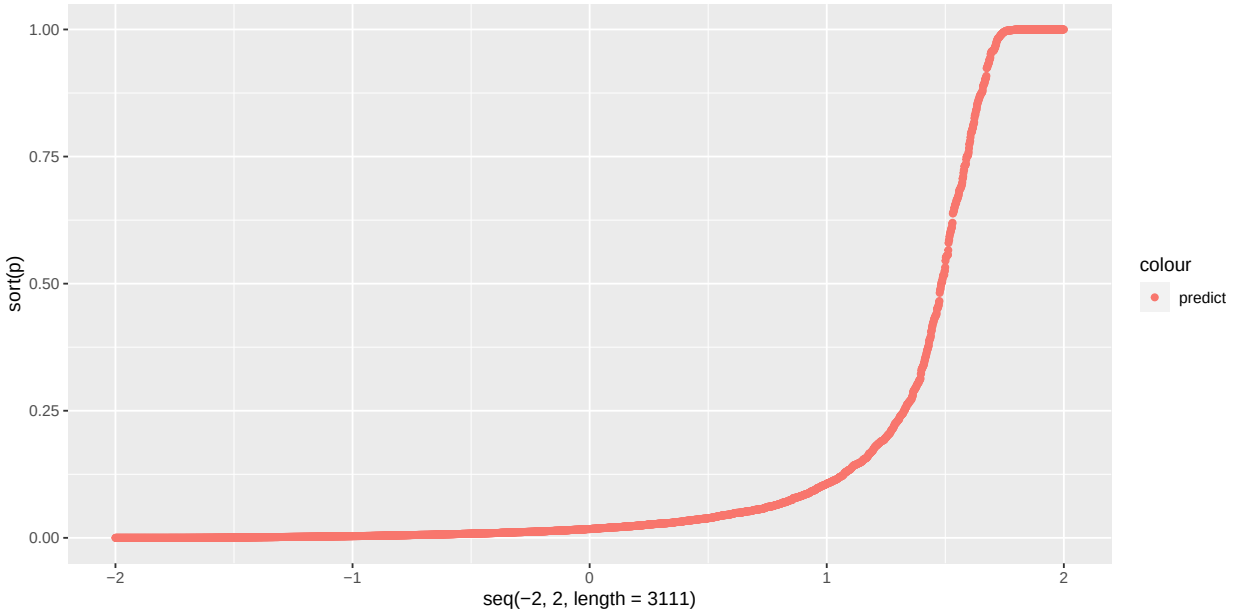

Figure 1: Fit Result of Initial Model

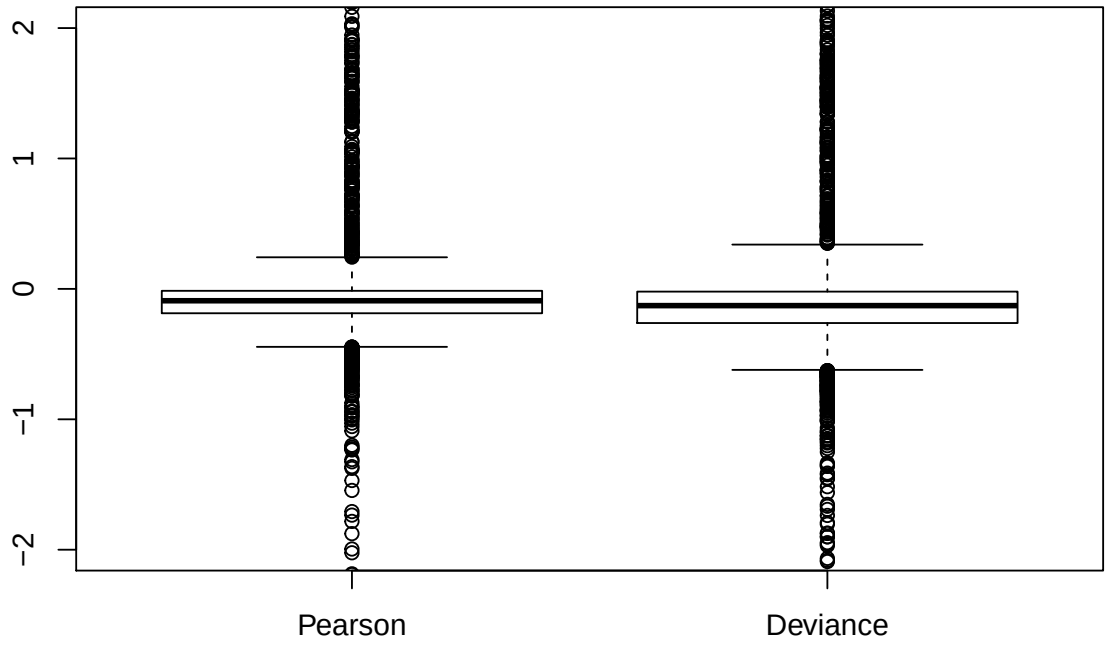

Figure 2: Boxplot of Deviance Residuals and Pearson Residuals

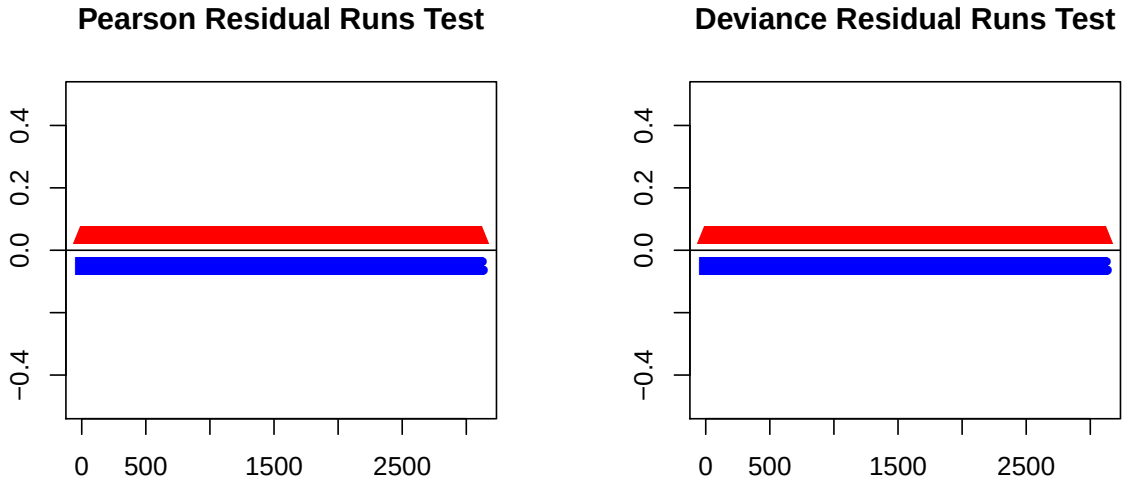

Figure 3: Runs Test Plots of Pearson Residuals and Deviance Residuals

Coefficients:	Estimate	Standard Error	z value	P- value
(Intercept)	2.581e+02	3.039e+02	0.849	0.3959
dem_2012	3.725e-04	2.547e-05	14.626	< 2e-16 ***
gop_2012	-6.446e-04	4.450e-05	-14.484	< 2e-16 ***
BachelorRate2016	1.498e+01	1.461e+00	10.251	< 2e-16 ***
MedianIncome2016	-1.699e-03	1.774e-04	-9.580	< 2e-16 ***
POP_ESTIMATE_2016	-9.869e-05	1.305e-05	-7.563	3.93e-14 ***
Economic_typology_20151	-7.426e-01	3.714e-01	-1.999	0.0456 *
Economic_typology_20152	4.741e-01	3.705e-01	1.280	0.2007
Economic_typology_20153	-1.859e-01	2.958e-01	-0.629	0.5296
Economic_typology_20154	2.274e-01	2.499e-01	0.910	0.3629
Economic_typology_20155	4.195e-01	2.821e-01	1.487	0.1369
N_POP_CHG_2016	4.158e+01	2.188e+01	1.900	0.0574 .
INTERNATIONAL_MIG_2016	1.404e+07	9.437e+06	1.488	0.1368
DOMESTIC_MIG_2016	1.404e+07	9.437e+06	1.488	0.1368
NET_MIG_2016	-1.404e+07	9.437e+06	-1.488	0.1368
GQ_ESTIMATES_2016	-4.169e-01	2.178e+00	-0.191	0.8482
Civilian_labor_force_2016	3.084e-04	3.144e-05	9.810	< 2e-16 ***
Unemployment_rate_2016	-2.301e+02	3.040e+02	-0.757	0.4491
Employed_2016	-2.620e+02	3.040e+02	-0.862	0.3888
Urban_influence_code_2013	-1.366e-01	3.169e-02	-4.310	1.63e-05 ***

Table 1: Fit Result of Initial Model

	Step	Df	Deviance Resid.	Df	Resid. Dev	AIC
1				3091	962.6990	1123.553
2	- Economic_typology_2015	5	10.52980720	3096	973.2289	1093.869
3	- GQ_ESTIMATES_2016	1	0.02164409	3097	973.2505	1085.848
4	- Unemployment_rate_2016	1	0.83604577	3098	974.0865	1078.642
5	- INTERNATIONAL_MIG_2016	1	2.03301566	3099	976.1196	1072.632
6	- DOMESTIC_MIG_2016	1	1.97413102	3100	978.0937	1066.563

	Step	Df	Deviance Resid.	Df	Resid. Dev	AIC
7	- N_POP_CHG_2016	1	1.17716156	3101	979.2708	1059.698
8	- NET_MIG_2016	1	0.37592364	3102	979.6468	1052.031

Table 2: BIC Result of Removed Variable

Variables in 2012 and 2016	Correlation
Bachelor Rate	0.98
Median Income	0.94
Population	1.00
Civilian Labor Force	1.00
Unemployment Rate	0.77

Table 3: Correlation between the independent variable which are collected in both 2012 and 2016

Type	Standardized Runs Statistic	p-value
Pearson Residual	-0.80692	0.4197
Deviance Residual	-0.80692	0.4197

Table 4: Run Test of Pearson Residuals and Deviance Residuals