An Analysis of the Presidential Elections

Wenchao Zhang

Abstract

In this project paper we illustrate how fundamental data can be used to make accurate forecasts of state-level presidential election outcomes. We analyzed 6 variables which may be related by the result of election, and find a good model for it.

Introduction

Fundamental models for forecasting elections are models that can make forecasts of the results of elections using only economic and political data available months before the election. Some people argue that Americans "vote their pocketbooks" in presidential elections. If this is the case, then the outcome of a presidential election should react the state of the economy immediately prior to that election. Alternatively, some people say that voters think mainly about how well the incumbent president has carried out his/her responsibilities in once. Therefore, presidential election outcomes should be related to public approval of the president in the months immediately preceding the election, and so on. [1] We now discuss such influence factors whether have actually related to vote rate.

Data description

The election data contains 8 variables, and each variable has 13 observations. The variables are: Year (from 1968-2012), Incumbent Vote, July Popularity, Peace, Future Problem, Lead Indicators, GNP Change, Second Term Indicator. Incumbent vote is the percentage of the two-party vote received by the candidate of the president's party. July popularity refers to the presidential popularity as measured by the Gallup poll in July before the election. Peace and prosperity is an index created by adding the percentage of two-party respondents who favored the incumbent party on keeping the United States out of war and on keeping the country prosperous (Gallup question). Future problem is the percentage of two-party respondents who favored the incumbent party on handling the country's most important problems (Gallup question). Leading indicators is the percentage change in the government's index of leading indicators during the first two quarters of the election year. It is set at zero if the change in one direction was not sustained for at least 3 months. GNP change is the non-annualized percentage change in GNP (constant dollars) from the fourth quarter of the year before the election to the second quarter of the election year. Second term is an indicator variable for party's second consecutive term in the White House. It is coded 1

if the party is heading into its second term and 0 otherwise. For the analysis of the data, I used the SPSS and Minitab.

Questions

Which is the best model for the prediction of Incumbent vote? Which variables should be chosen in the model? How can we have the best fit of the model?

Full model analysis

Assume that the full model is that

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_2 x_P + \beta_3 x_F + \beta_4 x_L + \beta_5 x_G + \beta_6 x_T + \epsilon_i$$

Which x_J is the value of JulPop, x_P is the value of Peace, x_F is the value of FutProb, x_L is the value of LeadIn, x_G is the value of GNPCha, x_T is the value of Term2.

We also assume that the error for every variable are normal distribution, and what's more, we have $E(\epsilon_i) = 0$, and $Var(\epsilon_i) = \sigma^2$.

After fitting by multiple regression models, we get that

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \\ \hat{\beta}_6 \end{pmatrix} = \begin{pmatrix} 29.66 \\ 0.1449 \\ 0.04638 \\ 0.1333 \\ -0.004 \\ 1.783 \\ 2.784 \end{pmatrix}$$

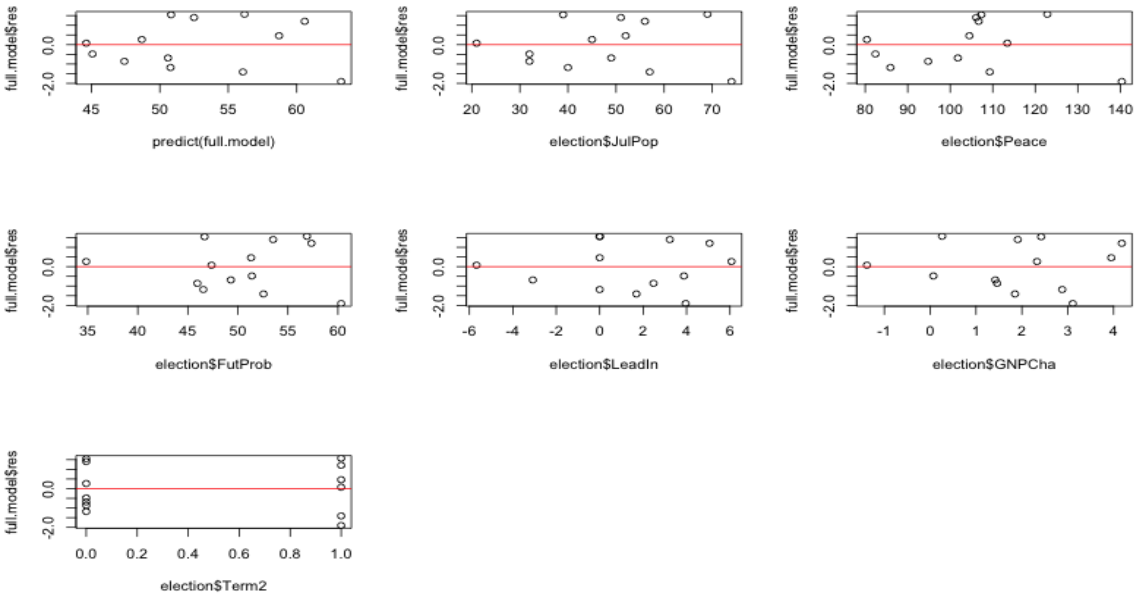
The P-values of $\hat{\beta}$ for corresponding variables are following:

$$p = \begin{pmatrix} 0.00109 \\ 0.04970 \\ 0.45864 \\ 0.30202 \\ 0.98477 \\ 0.00411 \\ 0.08199 \end{pmatrix}$$

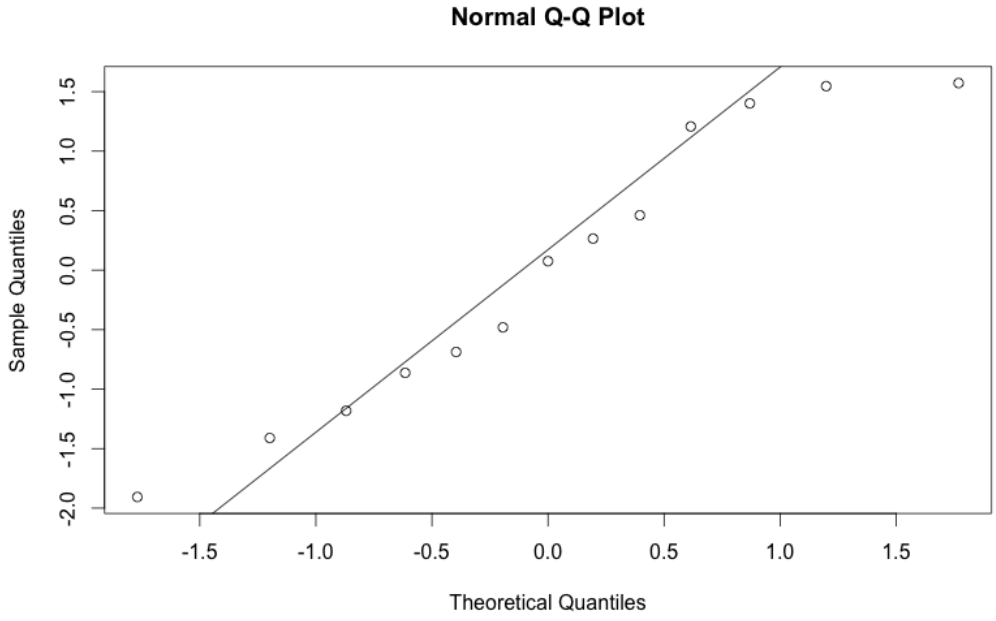
Hence, we can see that the Null Hypothesis that $\hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5 = 0$ are not rejected;
Therefore, we can consider for taking it out of model later.

Full model Residuals Diagnostics

First we check the constant variance assumption by looking at the variable versus residuals graph. However, we have only 13 observations; therefore, it is hard to check the constant variance assumption.



Then, residual normality check:



We can see that the data approximately follow normality, but it has a little bit tails off, and also there are only 13 observations; therefore, it is hard to determine the normality of the residuals.

Influential points or Outlier

First of all, is there any high leverage point, which affects the regression model a lot? By using the build in function in matlab, which provides the tr(H). Then, by the rule of thumb,

If $h_{ii} > \frac{2k}{n}$ then, the data should consider as a high leverage point; however, there is no high leverage or influential point in this data set.

Secondly, we check if there exists an outlier.
Using build-in function of finding the Studentized residuals. Then we can compare is to the Bonferroni values. If the Studentized residual is greater than Bonferroni value[2]; then it will be considered as an outlier.

For which Bonferroni Value is equal to student t-distribution, $n - p, \frac{\alpha}{2}$.

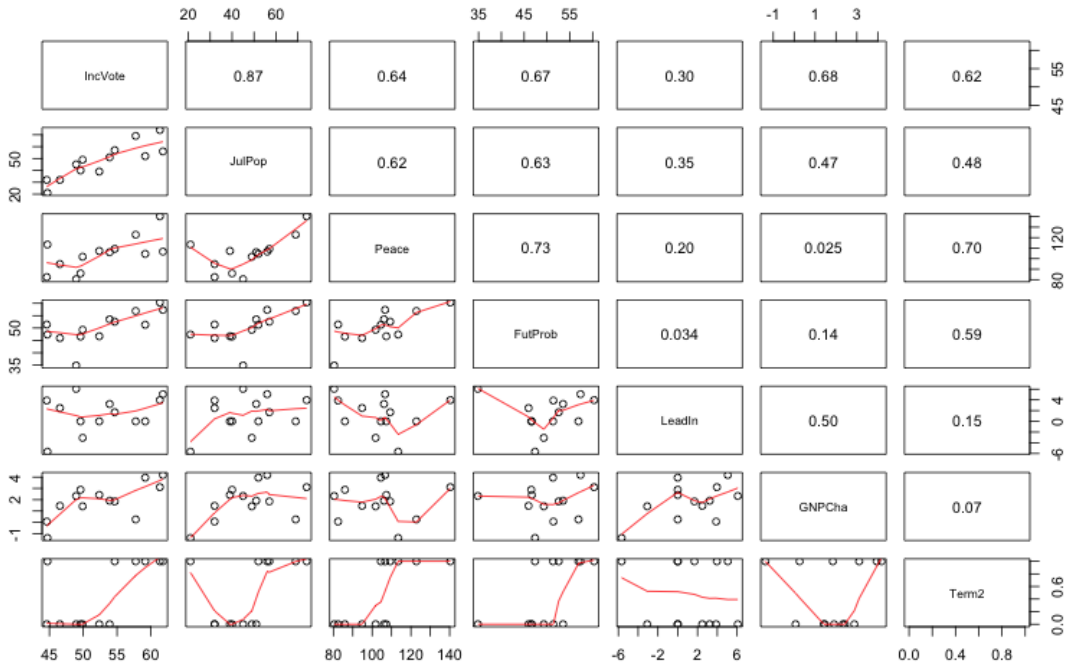
The Bonferroni Value is equal to student-t (0.05/26, with degree of freedom=6)
Which is equal to -5.076.
We cannot find any outlier in this model.

Serial Correlation of the data errors

As we see that the data is collected over years, so there may have some correlation between the successive errors.
Then we estimate the autocorrelation of the$(\epsilon_t, \epsilon_{t-1})$, where ϵ_{t-1} is the next observation error.
Then we get $corr(\epsilon_t, \epsilon_{t-1}) = 0.05228$, which is not big at all. It satisfies the assumption of independence in error.

Collinearity

Are there any relations between the variable?
By the correlation graph below, we can see that all the variables except LeadIn are highly correlated to the IncVote. Also, we can see that most of the variables are highly correlated to the others variables. Therefore, we can see that JulPop has the largest correlation with the IncVote, it seems that it is highly correlated to the IncVote, but we need to look forward into the model to find out.



Variable selection

Using the Akaike information criterion(AIC)[3], Bayesian information criterion(BIC)[4], and adjusted R-square to find a best-reduced model. When we start the model selection, we can start either forward elimination or backward elimination, or we can use the step function in R to find the best-reduced model.

In this data set, we have only 13 observations; therefore AIC may be a better criterion, but we still look at both AIC and BIC.

From above,

The P-value of $\hat{\beta} = \begin{pmatrix} 0.00109 \\ 0.04970 \\ 0.45864 \\ 0.30202 \\ 0.98477 \\ 0.00411 \\ 0.08199 \end{pmatrix}$

First, we use backward elimination, and we start from the full model, and eliminate the valuable with the highest p-value, and the second highest, and we keep eliminate and count the AIC, BIC, and adjusted R-square until the last variable. If the AIC or BIC become larger than the model without AIC; then we can eliminate this variables; moreover, we should also look at the adjusted R-square. If the adjusted R-square decreases a lot after the elimination, we do not eliminate this variable.

For the full model, we get the

$$AIC = 56.40956\,, BIC = 60.92916, adjusted\ R^2 = 0.9206$$

Then we first eliminate the LeadIn variable, and now the reduced model becomes

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_2 x_P + \beta_3 x_F + \beta_5 x_G + \beta_6 x_T + \epsilon_i$$

and the AIC, BIC, and adjusted R-square of the new reduced model is equal to

$$AIC = 54.41042\,, BIC = 58.36507, adjusted\ R - square = 0.9319$$

Therefore, LeadIn can be dropped from the full model because AIC and BIC is lower than full model, and adjusted R square increase compared to the full model.

Eliminate the Peace, the new reduce model becomes

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_5 x_G + \beta_6 x_T + \epsilon_i$$

and the AIC, BIC, and adjusted R-square of the new reduced model is equal to

$$AIC = 53.95319\,, BIC = 57.34288, adjusted\ R - square = 0.9329$$

Here again, we can see that AIC and BIC is lower than the full model, and the adjusted R-square increased; therefore, we can eliminate Peace from the full model.

Next, Eliminate the FutProb, the new reduce model becomes

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_5 x_G + \beta_6 x_T + \epsilon_i$$

and the AIC, BIC, and adjusted R-square of the new reduced model is equal to

$$AIC = 56.16344\,, BIC = 58.98819, adjusted\ R - square = 0.9176$$

We can see that AIC and BIC increase, but the adjusted R-square decrease.
Therefore, FutProb should not be dropped from the full model.

Then, eliminate the Term2, the new reduce model becomes

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_5 x_G + \epsilon_i$$

and the AIC, BIC, and adjusted R-square of the new reduced model is equal to

$$AIC = 61.84002\,, BIC = 64.66476, adjusted\ R - square = 0.8724$$

We can see that AIC and BIC increase, but the adjusted R-square decrease.
Therefore, Term2 should not be dropped from the full model.

Then, eliminate the JulPop, the new reduce model becomes

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_5 x_G + \epsilon_i$$

and the AIC, BIC, and adjusted R-square of the new reduced model is equal to

$$AIC = 64.42023\,, BIC = 67.24498, adjusted\ R - square = 0.8444$$

We can see that AIC and BIC increase, but the adjusted R-square decrease.
Therefore, JulPop should not be dropped from the full model.

Lastly, eliminate the GNPCha, the new reduce model becomes

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_6 x_T + \epsilon_i$$

and the AIC, BIC, and adjusted R-square of the new reduced model is equal to

$$AIC = 70.38085\,, BIC = 73.2056, adjusted\ R - square = 0.7539$$

We can see that AIC and BIC increase, but the adjusted R-square decrease.

Therefore, GNPCha should not be dropped from the full model.

To sum up, we have following table:

backward Elimination	1st drop LeadIn	2nd drop Peace	3rd drop FutProb	4th drop Term2	5th drop JulPop	6th drop GNPCha
AIC	54.4104	53.9532	56.1634	61.84	64.4202	70.3809
BIC	58.3651	57.3429	58.9882	64.6648	67.2450	73.2056
Adjusted R-Square	0.9319	0.9329	0.9176	0.8724	0.8444	0.7539
Keep/Drop Variable	Drop	Drop	Keep	Keep	Keep	Keep

Finally, by the backward elimination, the final reduced model that we will choose

$$\boldsymbol{IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_5 x_G + \beta_6 x_T + \epsilon_i}$$

Secondly, using the forward elimination, and we start with an empty model with no variable, and every time we add an variable with the lowest probability, and compare the AIC, BIC, and adjusted R-square to see keep the variable in model or not. The comparing method is same as backward elimination. We keep add variable and count the AIC, BIC, and adjusted R-square until the last variable. If the AIC or BIC become larger than the model without AIC; then we can add this variables; moreover, we should also look at the adjusted R-square. If the adjusted R-square decreases after adding variable, we should not add this variable in the model. And also we have following table similarly:

Forward Elimination	1st add GNPCha	2nd add JulPop	3rd add Term2	4th add FutProb	5th add Peace	6th add LeadIn
AIC	80.323	66.0492	56.1634	53.9532	54.4104	56.4096
BIC	82.018	68.3090	58.9882	57.3429	58.3651	60.9292
Adjusted R-Square	0.4115	0.8149	0.9176	0.9329	0.9319	0.9206
Add / Drop Variable	Add	Add	Add	Add	Drop	Drop

Therefore, By the Forward Elimination, we have the same reduced model as backward elimination, which is

$$\boldsymbol{IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_5 x_G + \beta_6 x_T + \epsilon_i}$$

Lastly, use step elimination to get a reduce model.
We still get the same reduced model

$$\boldsymbol{IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_5 x_G + \beta_6 x_T + \epsilon_i}$$

Therefore, we can see that the best reduced model should be

$$IncVote_i = \beta_0 + \beta_1 x_J + \beta_3 x_F + \beta_5 x_G + \beta_6 x_T + \epsilon_i$$

For which,

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \\ \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 31.6122 \\ 0.1624 \\ 0.1733 \\ 1.6761 \\ 3.2973 \end{pmatrix}$$

And for the P-value of the reduced-model coefficient

$$p = \begin{pmatrix} 5.86 \times 10^{-5} \\ 0.00711 \\ 0.11838 \\ 0.00105 \\ 0.01658 \end{pmatrix}$$

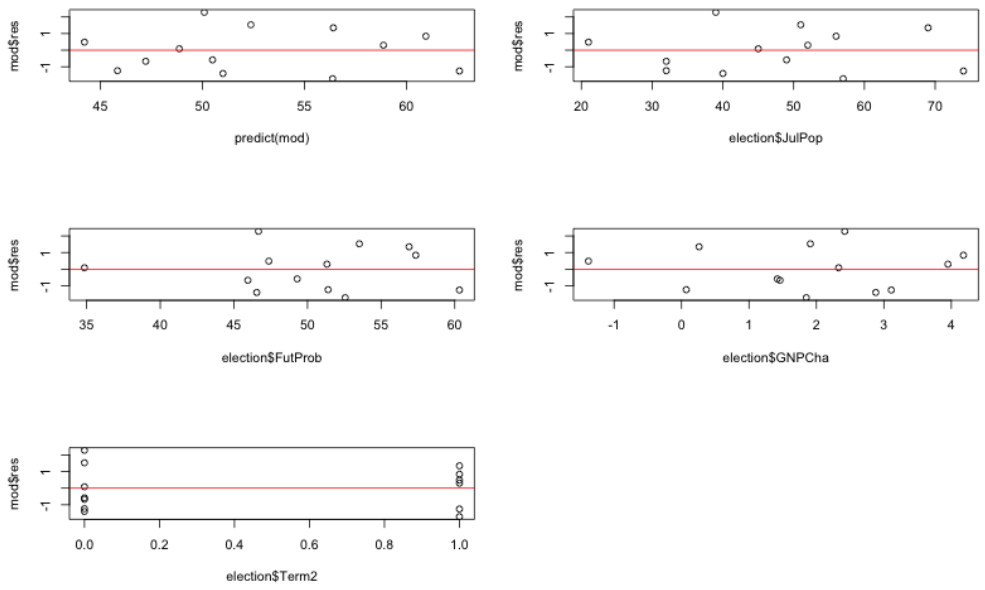
We can see that the β_0 is a highly significant, and β_1, β_5 are significant with 99.99% confidence interval. β_6 is also significant within 95%confidence interval.

We can see why we have this result because peace and is a question about the attitude of he incumbent party of war, but it may not be the most US citizens concern about, and also, Leading indicators may not be concern by the citizens also, or most of the citizens may not know this indicators also.

We can see that the July popularity is important for predicting the Incumbent vote because we can see that this is like a vote before the real vote; therefore, it is hard to change their decision within a few month, and that is why July popularity. Also, the FutProb is easy to see why it is highly related to incumbent vote because it is the favored of the incumbent on handling the country most important problems. It is no doubt that people will choose to vote an incumbent party if they can deal with the problem in their flavor way. At last, I think GNP change and Second Term is useful for predicting because if that is the second term of the incumbent, GNP change indicate how the incumbent party did good or bad on economics in his previous term, and how the incumbent party did in the economic is a good consideration for people to choose.

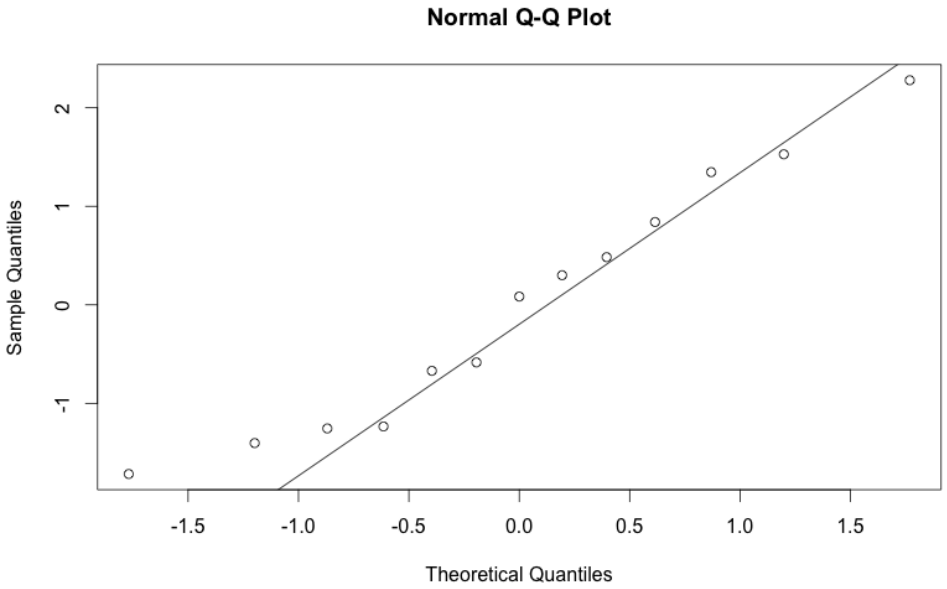
Reduced Model Residuals Diagnostics

Constant variance of residuals



We can see some kinds of constant variance pattern in the Left upper corner graph, and the last graph, but it is hard to say it is constant variance because the observation is not enough to tell it following constant variance or not.

Residuals normality



We can see the residual normality changes after reducing the model. It seems that the full model has a more normal residuals then the reduce one. Also, we can compare the graph and see that the

full model has a tail off at the large quartiles, but the reduced model tails off at the small quartiles, but overall it still can see the residuals is following the normal distribution, but it seems like t-distribution more.

Influential points of outliers in reduced model

First of all, is there any high leverage point, which affects the regression model a lot? By using the build in function in matlab, which provides the tr(H). By the rule of thumb,

If $h_{ii} > \frac{2k}{n}$ then, the data should considered as a high leverage point

We can find two influence points in the reduced model, which is point 8 and 9.
Then, check if any outlier is in this model. However, there is no high leverage or influential point in this data set as well.

Using build-in function of finding the Studentized residuals. Then we can compare is to the Bonferroni values. If the Studentized residual is greater than Bonferroni value; then it will be considered as an outlier.

For which Bonferroni Value is equal to student t-distribution,$n - p, \frac{\alpha}{2}$

The Bonferroni Value is equal to student- t(0.05/26, with degree of freedom=7)
Which is equal to -4.239.
We cannot find any outlier in this model.

Serial Correlation of the data errors in reduced model

Then we estimate the autocorrelation of the$(\epsilon_t, \epsilon_{t-1})$ in the reduced-model, where ϵ_{t-1} is the next observation error.
Then we get $corr(\epsilon_t, \epsilon_{t-1}) = -0.1888$, which is bigger then the full model
Therefore, it seems to be acceptable for a small sample size model.
It might not violate the independence assumption

Virtual Prediction using reduced Model

Assume that we don't have 1996 IncVote Value, and we want to predict the 1996 IncVote; therefore, we take out the value of 1996 in the data set.
And the model now is

$$IncVote_i^{(1996)} = \beta_0 + \beta_1 x_J{}^{(1996)} + \beta_3 x_F{}^{(1996)} + \beta_5 x_G{}^{(1996)} + \beta_6 x_T{}^{(1996)} + \epsilon_i$$

Now the $\hat{\beta} = \begin{pmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_3} \\ \widehat{\beta_5} \\ \widehat{\beta_6} \end{pmatrix} = \begin{pmatrix} 32.2797 \\ 0.17081 \\ 0.15314 \\ 1.63843 \\ 3.69298 \end{pmatrix}$

Then, we can try to predict 1996 IncVote value by this model.

$$IncVote_{i,1996} = \beta_0 + \beta_1 x_{J,1996} + \beta_3 x_{F,1996} + \beta_5 x_{G,1996} + \beta_6 x_{T,1996} + \epsilon_i$$

We get that $IncVote_{i,1996} = 56.79064$

By the predict function from R, we have the 95% confidence interval for $\hat{Y}$

$\hat{Y} \pm s \times t_{n-k,\frac{\alpha}{2}}$, which s is the standard error=0.7312829, and have DF=7

we have that $56.79064 \pm 0.7312829 \times 2.36$

Confidence interval for $\hat{Y} = \{55.06143 \leq \hat{Y} \leq 58.51985\}$

However, we can see that the true value of 1996 IncVote equal to 54.66 which is not inside the 95% confidence interval.

When we look further into the reduced model, we can see that the R-square is equal to 0.9553,which means that there is only 4.4% residuals is not explained by the model, which is a pretty good fit, and we can also see that except for FutProb; all others variable was not rejected by the Null Hypothesis that $\beta_0, \beta_1, \beta_5, \beta_6 = 0$, in 95% confidence interval; therefore, I think this model is good, but we can see that FutProb has no significant level even in 90% confidence interval; therefore, we see that 1996 is not inside its confidence interval, it might be a problem caused by FutProb. Also, from previous plot, we can see that point 13, which is the data of 1996, has one of the largest residual; therefore, it is not hard to see why it is not inside its confidence interval.

Summary

It can be seen that Peace and LeadIn should be dropped from the full model because it increase the AIC and decrease the adjusted R-square. Also, it is a pretty good model because we have a large R-square for the reduced-model, which means most of the residuals are explained by the model, but we have only 13 observations; it is hard to confirm all the assumption of regression model. Therefore, the model may not be accurate enough to predict the future value, but it is good enough for predicting.

Reference

[1] Ray C. Fair, Presidential and Congressional Vote-Share Equations: November 2010 Update

[2] Audic, S. and J. M. Claverie (1997). The significance of digital gene expression profiles. Genome Res 7(10): 986-95.

[3] Akaike, Hirotsugu. A new look at the statistical model identification. IEEE Transactions on

Automatic Control. 1974, **19** (6): 716–723.
[4] Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. WIREs Computational Statistics4, 199.203.