Analysis of 2016 Presidential Election Data

Yi-Chen Zhang
Department of Statistics and Probability, Michigan State University, USA.
E-mail: zhang318@stt.msu.edu

Abstract

In the report, we perform logistic regression to analyze United States presidential election data. We split the election polls data set collected in 2016 into two blocks: a training and a test set. We use the training set to construct several models, then evaluate and diagnose models to choose the best model. The model adequacy is also performed. The best model is then fitted to the test set. The results of classification of test data show a satisfactory performance under the best model.
Keywords: Logistic regression model, presidential election.

1 Introduction

Since the beginning of the 20th century, there has been a great interest in the prediction of the outcome of presidential elections. To make a good prediction, a lot of election polls are conducted during the election period. During the 2016 presidential election period, the website fivethirtyeight.com conducted more than ten thousand polls at various locations and times, and using various sampling methods. An effective analysis of the resulting data set consititutes a big challenge for statisticians.

For simplicity, in our analysis of 2016 presidential election data set, we start our applications of logistic regression by considering only the two major political parties: Democratic party (Dem) and Republican party (Rep). Logistic regression is a technique that is well suited for examining the relationship between a categorical response variable and one or more categorical or continuous predictor variables.

The report is organized as follows: In section 2, we introduce the presidential election data and factors that are affecting the prediction accuracy of election polls. The model fitting, model selection and how to interpret the model are discussed in section 3. Section 4 provides the prediction accuracy on the test data based on the model we build from the training data set. The further analysis of this data set is discussed in Section 5.

2 The presidential election data

We create several variables based on the factors which might affect the prediction accuracy of election polls: Margin (the difference between the percentage of Democratic and Republican voters), Survey Date (the date the poll was conducted), and Pollster(the organization that conducted the poll). This idea is adopted from Lab 3. Table 1 shows the distribution of 2016 presidential election polls survey for each state, classified by political party (Democratic - Dem and Republican - Rep), Margin, Survey Date, and Pollster.

To make a further inference about election polls we also downloaded the 2016 presidential election results from wikipedia. We consider the case where the response y_i is binary, assuming only two values which, for convenience, we code as one or zero. We define

$$y_i = \begin{cases} 0, & \text{if the Democratic party wins the state} \\ 1, & \text{if the Republican party wins the state.} \end{cases}$$

Table 1: 2016 presidential election polls

States	Dem	Rep	Margin	Survey Date	Pollster
AK	41.00	44.00	-3.00	Nov 07 2016	Gravis Marketing
⋮	⋮	⋮		⋮	⋮
AK	31.00	48.00	-17.00	Sep 09 2016	Google Consumer Surveys
AL	36.00	55.00	-19.00	Nov 08 2016	SurveyMonkey
⋮	⋮	⋮		⋮	⋮
AL	38.30	53.84	-15.54	Oct 15 2016	IPSOS
⋮	⋮	⋮		⋮	⋮
WY	21.00	60.00	-39.00	Nov 08 2016	SurveyMonkey
⋮	⋮	⋮		⋮	⋮
WY	21.98	40.77	-18.79	Sep 09 2016	Google Consumer Surveys

We view y_i as a realization of a random variable Y_i that can take the values one and zero with probabilities π_i and $1 - \pi_i$, respectively.

In our analysis of these data we will view the y_i as the response or dependent variable of interest and State, Margin, Lagtime and Pollster as predictors. Note that the Margin is calculated by Dem subtracted from Rep, and the Lagtime is the number of days between the polling date and the presidential election data of 2016. In this analysis, State and Pollster are treated as discrete factors, Margin and Lagtime are treated as continuous variables.

3 Model fitting and model selection

3.1 Model fitting

Since we assume only one of two parties will win the state, the election result is a binary outcome for each state. To exam the relationship between a categorical response variable and more predictor variables, the logistic regression is best used in this condition.

We first use the contrasts() function in R to set up the categorical variable States and Pollers. We see that the state "AK" will be used as the reference state, and the poller "CNN/Opinion Research Corp." will be used as the reference poller in our analysis. This contrasts step often is crucial for obtaining correct estimate of the coefficients in logistic regression model.

We then split the data into two chunks: training and test set. The training set will be used to fit our model which we will be testing over the test set. The logistic regression model can be easily fit by the glm() function in R and the logit link is the default link function for binomial model.

```
> logit_full <- glm(resp~states+margins+lagtime+pollers, data=train,
+ family=binomial(link="logit"))
```

One can use the summary() function to obtain the summary of the fitted model and the anova() function to analyze the table of deviance.

3.2 Model selection

We then proceed to perform the model selection and to get a parsimonious model. First, a Wald type test is performed to test the hypothesis that the coefficient of a predictor variable in the model is significantly different from zero. If the test fails to reject the null hypothesis, this

Table 2: Analysis of Deviance Table

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			2136	2241.5		
states	50	1100.81	2086	1140.7	< 2.2e-16	***
margins	1	48.06	2085	1092.6	4.141e-12	***
lagtime	1	0.35	2084	1092.3	0.552467	
pollers	24	50.42	2060	1041.8	0.001253	**

suggests that removing the variable from the model will not substantially harm the fit of that model. We can see from the summary result listed in the appendix, the Wald type test shows that the Lagtime is not statistically significant. As for the statistically significant variables, Margin has the lowest p-value suggesting a strong association of the States and Pollster with the probability of having voted for Democratic.

The ANOVA is reported in Table 2. The difference between the null deviance and residual deviance shows how our model is doing against the null model (a model with only the intercept). The wider this gap, the better. Analyzing the table, we can see the drop in deviance when adding each variable sequentially one at a time. Again, adding State, Margin, and Poller significantly reduces the residual deviance. A large p-value here indicates that the model without the variable explains the less amount of variation. Ultimately what we would like to see is a significant drop in deviance.

The R function `step()` can be used to perform variable selection as well, and the model selection criterion is based on the Akaike information criterion (AIC), which is a measure of the relative quality of statistical models for a given set of data. We ran "forward selection", "backward elimination", and "both direction" in `step()` function and those procedures give rise to results that are equivalent to the result in the ANOVA table.

3.3 Likelihood Ratio Test

A logistic regression is said to provide a better fit to the data if it demonstrates an improvement over a model with fewer predictors. This is performed using the likelihood ratio test, which compares the likelihood of the data under the full model against the likelihood of the data under the best model. Given that H_0 holds that the reduced model is true, removing predictor variables from a full model will almost always make the model fit less well, which provide evidence against the reduced model in favor of the current model compared the full (with Lagtime) against best model (without Lagtime) using the `anova()` function in R. The result also suggests Lagtime is not a significant variable. We also ran `anova()` function to compare the best model with the model dropped one variable out. These results are listed in the Appendix and indicating the best model is the model with variable State, Margin, and Pollster.

```
> anova(logit_reg, logit_full, test="Chisq")
Analysis of Deviance Table

Model 1: resp ~ statesFAC + margins + pollersFAC
Model 2: resp ~ statesFAC + margins + lagtime + pollersFAC
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2061     1041.9
2      2060     1041.8  1 0.025096   0.8741
```

3.4 Interpreting the results of our logistic regression model

We first formulate the best model derived from the model selection result:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \texttt{states} + \beta_2 \texttt{margins} + \beta_3 \texttt{pollers},$$

and the following is some part of the summary output. The full summary output can be found in the Appendix.

```
Call:
glm(formula = resp ~ states + margins + pollers,
 family = binomial(link = "logit"), data = train)

Deviance Residuals:
  Min        1Q    Median        3Q       Max
-2.75040   0.00003   0.00012   0.32932   2.89283

Coefficients:
                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)       0.40390     0.87452    0.462   0.64418
statesAL         16.55018  1758.64511    0.009   0.99249
...
statesWY         15.98957  2041.22655    0.008   0.99375
margins           0.14614     0.02329    6.276  3.48e-10 ***
pollersDanJones  -0.80693     1.38404   -0.583   0.55988
...
pollersYouGov    -0.45006     0.74565   -0.604   0.54612
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2241.5  on 2136  degrees of freedom
Residual deviance: 1041.9  on 2061  degrees of freedom
AIC: 1193.9

Number of Fisher Scoring iterations: 18
```

As for the statistically significant variables, the margin has the lowest p-value suggesting a strong association of the margins of the state with the response variable. The positive coefficient for this predictor suggest that all other variables being equal, a unit increase in margin increases the log odds voting for Democratic by 0.1464. Since state is a dummy variable and the baseline is the state "AK", the coefficient statesAL in the table means that being in state "AL" increase the log odds voting for Democratic by 16.55 compared with the baseline state "AK". Similar interpretation for Pollster variable, conducting poll by poller DanJones decreases the log odds by 0.8069 compared with the baseline poller "CNN/Opinion Research Corp."

3.5 Residual

The deviance residual plot and Pearson residual plot are shown in Figure 1(a) and Figure 1(b). If the fitted logistic regression model was true, we would expect to see a horizontal band with most of the residuals falling withing ± 3. Figure 1(b) the Pearson residual plot indicates some potential outliers present in our data set. With purely binary response, the usual residual diagnositic plots are not very useful. The two parallel curves in these two residual plots are caused by the fact that the response has only two possible values, but the predicted values are continuous. If you just consider the 0s, then you get one of those curves of points as the predictor varies over the full range. Likewise, if you consider the 1s, you get the other curve of points. In other words, this graph does not tell us if the model is appropriate or not.

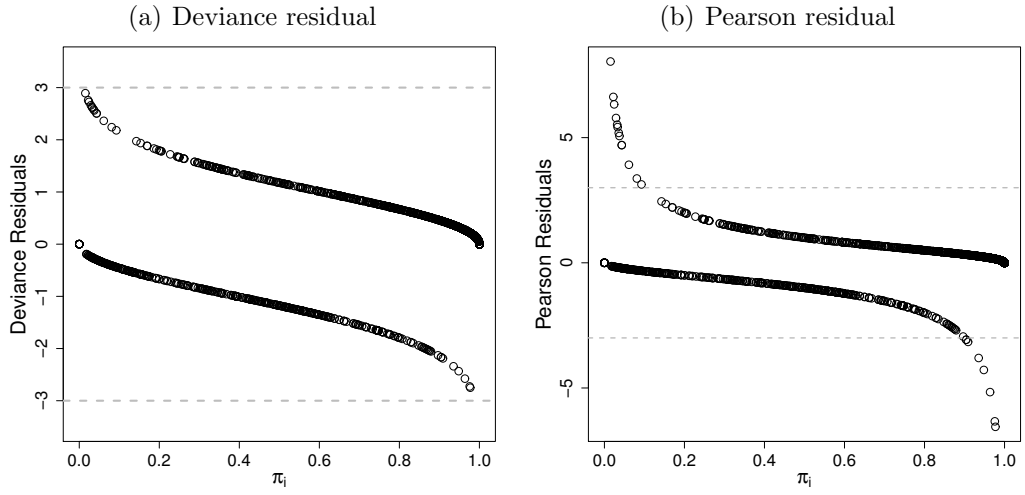

Figure 1: Deviance residual and Pearson residual plot.

4 Prediction of the model

4.1 Assessing the predictive ability of the model

In the steps above, we briefly evaluated the fitting of the model, now we would like to see how the model is doing when predicting y on a new set of data. By setting the parameter `type='response'`, R will output probabilities in the form of $P(y = 0|X)$. Our decision boundary will be 0.5. If $P(y = 0|X) > 0.5$ then $y = 0$ otherwise $y = 1$.

```
> fitted.results <- predict(logit_reg,newdata=test,type='response')
> fitted.results <- ifelse(fitted.results > 0.5,1,0)
> misClasificError <- mean(fitted.results != test$resp)
> print(paste('Accuracy',1-misClasificError))
[1] "Accuracy 0.885981308411215"
```

The 0.89 accuracy on the test set is quite a good result. However, this number can be vary and this result is somewhat dependent on how do we split data.

4.2 ROC Curve

In this step, we are going to plot the ROC curve and calculate the AUC (area under the curve) which are typical performance measurements for a binary classifier. The ROC is a curve generated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings while the AUC is the area under the ROC curve. As a rule of thumb, a model with good predictive ability should have an AUC closer to than to 0.5.

```
> p <- predict(logit_reg, newdata=test, type="response")
> pr <- prediction(p, test$resp)
> prf <- performance(pr, measure = "tpr", x.measure = "fpr")
> auc <- performance(pr, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.9557519
```

The AUC of 0.9558 indicates a good performance of the best model.

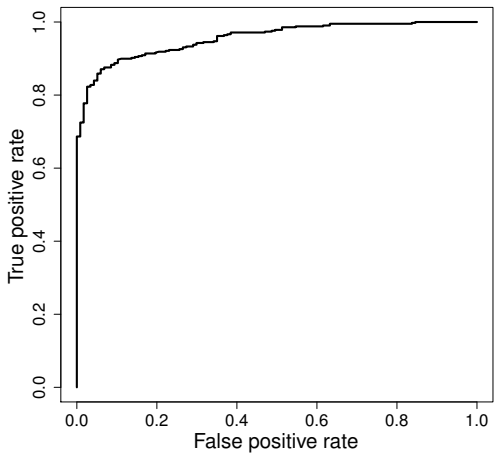

Figure 2: ROC

4.3 K-Fold Cross Validation

In the last step, we ran 10-fold cross-validation, which partitions the data into 10 equally sized segments. One-fold is held out for validation while the other nine folds are used to train the model and then used to predict the target variable in our test data.

```
> library('caret')
> ctrl <- trainControl(method="repeatedcv",number=10,savePredictions=TRUE)
> mod_fit <- train(as.factor(resp)~states+margins+pollers, data=train,
+ method="glm",family=binomial(link="logit"),trControl=ctrl)
> pred <- predict(mod_fit, newdata=test)
> confusionMatrix(data=pred, test$resp)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  84  28
         1  33 390

               Accuracy : 0.886
                 95% CI : (0.856, 0.9117)
    No Information Rate : 0.7813
    P-Value [Acc > NIR] : 2.333e-10

                  Kappa : 0.6611
 Mcnemar's Test P-Value : 0.6085

            Sensitivity : 0.7179
            Specificity : 0.9330
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.9220
             Prevalence : 0.2187
         Detection Rate : 0.1570
   Detection Prevalence : 0.2093
      Balanced Accuracy : 0.8255
```

```
        'Positive' Class : 0
```

The results of 10-fold cross-validation also indicates a satisfactory classification accuracy rate.

5 Further analysis

Here is some additional work we can do for further analyzing this data. First, we can try to change the link function. Changing the link function will change the interpretation of the coefficients entirely; the β_j's will no longer be log-odds ratios. Secondly, we can assume heterogeneity of variance between each states. In our analysis, we assume the variance is a constant all over the state, which could lead to a overdispersion or underdispersion problem in our model.

6 Appendix

The summary table for the full model.

```
-------------------------------------------------------------------------------
  Call:
  glm(formula = resp ~ statesFAC + margins + lagtime + pollersFAC,
      family = binomial(link = "logit"), data = train)

  Deviance Residuals:
       Min         1Q     Median         3Q        Max
  -2.74924   0.00003   0.00012   0.32836   2.89050

  Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
  (Intercept)                       4.230e-01  8.829e-01   0.479  0.63191
  statesAL                          1.655e+01  1.759e+03   0.009  0.99249
  statesAR                          2.172e+00  1.171e+00   1.855  0.06364 .
  statesAZ                          1.413e-01  6.403e-01   0.221  0.82536
  statesCA                          1.625e+01  1.561e+03   0.010  0.99169
  statesCO                          5.999e-01  6.508e-01   0.922  0.35660
  statesCT                          1.790e+01  1.785e+03   0.010  0.99200
  statesDC                          1.559e+01  1.567e+03   0.010  0.99207
  statesDE                          1.816e+01  2.008e+03   0.009  0.99278
  statesFL                         -1.374e+00  6.001e-01  -2.290  0.02204 *
  statesGA                          2.687e+00  9.519e-01   2.822  0.00477 **
  statesHI                          1.669e+01  2.059e+03   0.008  0.99353
  statesIA                         -5.429e-01  6.181e-01  -0.878  0.37976
  statesID                          1.727e+01  1.578e+03   0.011  0.99127
  statesIL                          1.777e+01  1.704e+03   0.010  0.99168
  statesIN                          6.730e-01  7.201e-01   0.935  0.34997
  statesKS                         -1.008e+00  6.654e-01  -1.514  0.13001
  statesKY                          1.740e+00  9.254e-01   1.880  0.06008 .
  statesLA                          1.781e+01  1.673e+03   0.011  0.99151
  statesMA                          1.617e+01  1.642e+03   0.010  0.99214
  statesMD                          1.615e+01  1.832e+03   0.009  0.99297
  statesME                          1.478e+00  8.239e-01   1.793  0.07290 .
  statesMI                         -3.739e+00  9.208e-01  -4.060 4.90e-05 ***
  statesMN                          2.595e+00  1.157e+00   2.243  0.02492 *
```

```
statesMO                          7.546e-02  6.538e-01   0.115  0.90812
statesMS                          1.755e+01  1.735e+03   0.010  0.99193
statesMT                          1.247e+00  8.358e-01   1.492  0.13568
statesNC                         -1.199e+00  5.964e-01  -2.010  0.04440 *
statesND                          1.659e+01  2.203e+03   0.008  0.99399
statesNE                          1.531e+00  1.169e+00   1.309  0.19054
statesNH                          1.418e+00  7.569e-01   1.874  0.06096 .
statesNJ                          2.330e+00  1.164e+00   2.001  0.04537 *
statesNM                          1.680e+00  9.307e-01   1.805  0.07105 .
statesNV                         -4.000e-01  6.094e-01  -0.656  0.51158
statesNY                          1.703e+01  1.446e+03   0.012  0.99060
statesOH                         -3.545e-01  6.052e-01  -0.586  0.55810
statesOK                          1.711e+01  1.739e+03   0.010  0.99215
statesOR                          1.789e+01  1.663e+03   0.011  0.99142
statesPA                         -3.044e+00  6.608e-01  -4.607 4.09e-06 ***
statesRI                          1.780e+01  2.099e+03   0.008  0.99323
statesSC                          2.693e+00  1.151e+00   2.340  0.01930 *
statesSD                          3.032e-01  8.831e-01   0.343  0.73132
statesTN                          1.766e+01  1.924e+03   0.009  0.99268
statesTX                          1.821e+01  1.609e+03   0.011  0.99097
statesUT                          1.485e+00  7.591e-01   1.957  0.05039 .
statesVA                          1.918e+00  9.094e-01   2.109  0.03496 *
statesVT                          1.566e+01  2.154e+03   0.007  0.99420
statesWA                          1.758e+01  1.815e+03   0.010  0.99227
statesWI                         -4.070e+00  8.143e-01  -4.999 5.77e-07 ***
statesWV                          1.393e+00  1.196e+00   1.165  0.24401
statesWY                          1.599e+01  2.041e+03   0.008  0.99375
margins                           1.462e-01  2.329e-02   6.276 3.47e-10 ***
lagtime                          -3.119e-04  1.969e-03  -0.158  0.87411
pollersDanJones                  -7.948e-01  1.387e+00  -0.573  0.56668
pollersEmerson College            2.182e-01  7.850e-01   0.278  0.78105
pollersGoogle Consumer Surveys   -8.422e-01  6.872e-01  -1.225  0.22039
pollersGravis Marketing          -1.066e+00  7.830e-01  -1.361  0.17337
pollersGreenberg Quinlan Rosner  -1.423e+00  9.732e-01  -1.463  0.14360
pollersInsiderAdvantage          -1.542e+00  1.419e+00  -1.087  0.27709
pollersIPSOS                     -2.110e-01  6.891e-01  -0.306  0.75946
pollersMaristColl                -8.896e-01  8.073e-01  -1.102  0.27051
pollersMarquette University      -1.715e+01  4.376e+03  -0.004  0.99687
pollersMasonDixon                -7.740e-01  1.424e+00  -0.543  0.58679
pollersMitchellResearch          -1.739e+01  3.549e+03  -0.005  0.99609
pollersMonmouthU                 -1.387e-01  8.387e-01  -0.165  0.86866
pollersPPP                        2.166e-01  7.427e-01   0.292  0.77058
pollersQuinnipiacU                3.330e-01  7.795e-01   0.427  0.66925
pollersRasmussen                  5.148e-02  7.331e-01   0.070  0.94401
pollersRemington                  9.750e-01  8.702e-01   1.120  0.26250
pollersSienaColl                  6.867e-01  1.083e+00   0.634  0.52593
pollersSuffolkU                   1.250e+00  1.048e+00   1.193  0.23305
pollersSurveyMonkey               6.840e-02  6.960e-01   0.098  0.92171
pollersSurveyUSA                  7.503e-01  9.388e-01   0.799  0.42414
pollersThe Times-Picayune        -2.292e+00  1.347e+00  -1.701  0.08898 .
pollersTrafalgar Group            1.092e+00  1.324e+00   0.824  0.40966
pollersUofNewHampshire           -1.067e-01  1.311e+00  -0.081  0.93512
pollersYouGov                    -4.579e-01  7.476e-01  -0.612  0.54021
```

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2241.5  on 2136  degrees of freedom
Residual deviance: 1041.8  on 2060  degrees of freedom
AIC: 1195.8

Number of Fisher Scoring iterations: 18
```

The summary table for the best model.

```
Call:
glm(formula = resp ~ statesFAC + margins + pollersFAC,
family = binomial(link = "logit"),
data = train)

Deviance Residuals:
Min        1Q    Median        3Q       Max
-2.75040   0.00003   0.00012   0.32932   2.89283

Coefficients:
                                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)                        0.40390    0.87452    0.462  0.64418
statesAL                          16.55018 1758.64511    0.009  0.99249
statesAR                           2.16896    1.17116    1.852  0.06403 .
statesAZ                           0.14036    0.64039    0.219  0.82651
statesCA                          16.24154 1561.61906    0.010  0.99170
statesCO                           0.59779    0.65076    0.919  0.35830
statesCT                          17.89896 1785.20612    0.010  0.99200
statesDC                          15.57879 1567.49220    0.010  0.99207
statesDE                          18.15506 2008.20172    0.009  0.99279
statesFL                          -1.37550    0.60016   -2.292  0.02191 *
statesGA                           2.68476    0.95172    2.821  0.00479 **
statesHI                          16.69268 2058.87989    0.008  0.99353
statesIA                          -0.54888    0.61710   -0.889  0.37376
statesID                          17.26669 1577.63943    0.011  0.99127
statesIL                          17.76810 1704.02277    0.010  0.99168
statesIN                           0.67128    0.72005    0.932  0.35119
statesKS                          -1.00760    0.66555   -1.514  0.13004
statesKY                           1.73658    0.92526    1.877  0.06054 .
statesLA                          17.80656 1672.77534    0.011  0.99151
statesMA                          16.16905 1642.53873    0.010  0.99215
statesMD                          16.15333 1832.36605    0.009  0.99297
statesME                           1.47393    0.82364    1.790  0.07353 .
statesMI                          -3.74368    0.92104   -4.065 4.81e-05 ***
statesMN                           2.59507    1.15741    2.242  0.02495 *
statesMO                           0.07457    0.65390    0.114  0.90921
statesMS                          17.54241 1735.96563    0.010  0.99194
statesMT                           1.24441    0.83575    1.489  0.13649
statesNC                          -1.20112    0.59636   -2.014  0.04400 *
```

```
statesND                          16.59154 2202.37339    0.008  0.99399
statesNE                           1.52744    1.16911    1.307  0.19138
statesNH                           1.41927    0.75692    1.875  0.06078 .
statesNJ                           2.32483    1.16381    1.998  0.04576 *
statesNM                           1.67925    0.93077    1.804  0.07121 .
statesNV                          -0.40037    0.60950   -0.657  0.51125
statesNY                          17.02636 1445.44832    0.012  0.99060
statesOH                          -0.35767    0.60506   -0.591  0.55443
statesOK                          17.10693 1738.77580    0.010  0.99215
statesOR                          17.88679 1663.05827    0.011  0.99142
statesPA                          -3.04574    0.66094   -4.608 4.06e-06 ***
statesRI                          17.79692 2098.80515    0.008  0.99323
statesSC                           2.69060    1.15114    2.337  0.01942 *
statesSD                           0.29924    0.88263    0.339  0.73459
statesTN                          17.65768 1923.65792    0.009  0.99268
statesTX                          18.21086 1609.47266    0.011  0.99097
statesUT                           1.48208    0.75886    1.953  0.05082 .
statesVA                           1.91737    0.90946    2.108  0.03501 *
statesVT                          15.65505 2153.70231    0.007  0.99420
statesWA                          17.58140 1815.18724    0.010  0.99227
statesWI                          -4.07416    0.81417   -5.004 5.61e-07 ***
statesWV                           1.38790    1.19524    1.161  0.24557
statesWY                          15.98957 2041.22655    0.008  0.99375
margins                            0.14614    0.02329    6.276 3.48e-10 ***
pollersDanJones                   -0.80693    1.38404   -0.583  0.55988
pollersEmerson College             0.22940    0.78137    0.294  0.76907
pollersGoogle Consumer Surveys    -0.83204    0.68396   -1.217  0.22379
pollersGravis Marketing           -1.05767    0.78095   -1.354  0.17563
pollersGreenberg Quinlan Rosner   -1.43675    0.96985   -1.481  0.13850
pollersInsiderAdvantage           -1.54176    1.42155   -1.085  0.27812
pollersIPSOS                      -0.20106    0.68602   -0.293  0.76946
pollersMaristColl                 -0.89698    0.80606   -1.113  0.26580
pollersMarquette University      -17.18544 4375.16066   -0.004  0.99687
pollersMasonDixon                 -0.77341    1.42179   -0.544  0.58646
pollersMitchellResearch          -17.37758 3548.72718   -0.005  0.99609
pollersMonmouthU                  -0.13541    0.83818   -0.162  0.87166
pollersPPP                         0.20353    0.73788    0.276  0.78268
pollersQuinnipiacU                 0.33256    0.77927    0.427  0.66956
pollersRasmussen                   0.06765    0.72570    0.093  0.92573
pollersRemington                   0.99088    0.86443    1.146  0.25168
pollersSienaColl                   0.70061    1.07898    0.649  0.51612
pollersSuffolkU                    1.25107    1.04798    1.194  0.23256
pollersSurveyMonkey                0.08607    0.68668    0.125  0.90025
pollersSurveyUSA                   0.74215    0.93651    0.792  0.42809
pollersThe Times-Picayune         -2.27392    1.34261   -1.694  0.09033 .
pollersTrafalgar Group             1.11174    1.31812    0.843  0.39899
pollersUofNewHampshire            -0.10498    1.31045   -0.080  0.93615
pollersYouGov                     -0.45006    0.74565   -0.604  0.54612
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)
```

```
  Null deviance: 2241.5  on 2136  degrees of freedom
  Residual deviance: 1041.9  on 2061  degrees of freedom
  AIC: 1193.9

  Number of Fisher Scoring iterations: 18
--------------------------------------------------------------------------------
```

The ANOVA comparison between the best model and the model with one variable dropped out.

```
> logit_reg <- glm(resp~states+margins+pollers, data=train, family=binomial(link="logit"))
> logit_reg1 <- glm(resp~margins+pollers, data=train, family=binomial(link="logit"))
> anova(logit_reg1, logit_reg, test="Chisq")
Analysis of Deviance Table

Model 1: resp ~ margins + pollers
Model 2: resp ~ states + margins + pollers
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      2111     1636.8
2      2061     1041.9 50   594.92 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> logit_reg2 <- glm(resp~states+pollers, data=train, family=binomial(link="logit"))
> anova(logit_reg2, logit_reg, test="Chisq")
Analysis of Deviance Table

Model 1: resp ~ states + pollers
Model 2: resp ~ states + margins + pollers
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      2062     1088.2
2      2061     1041.9  1   46.341 9.938e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> logit_reg3 <- glm(resp~states+margins, data=train, family=binomial(link="logit"))
> anova(logit_reg3, logit_reg, test="Chisq")
Analysis of Deviance Table

Model 1: resp ~ states + margins
Model 2: resp ~ states + margins + pollers
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      2085     1092.6
2      2061     1041.9 24   50.744 0.001138 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```