Illapani_Project_IS606

Srini Illapani

December 6, 2015

Introduction:

Do number of endorsements increase the chances of winning the primaries and party nomination?
We would also attempt to see if the money raised by the candidates is any indicator of winning the primaries, and compare the two variables as to which one correlates more to obtaining the majority vote share during the primaries.

We will be looking at the democratic and republican party candidates for presidential primaries from the year 1980 to 2012.

The presidential candidate endorsements for primary elections is an observational study performed by FiveThirtyEight by collecting the data as mentioned above.

Data:

Fundraising data since 1992 is taken from the Federal Election Commission website. Data from 1980 to 1988 is from various news articles at the time of the filing deadline. This data is preliminary, though rarely differs greatly from finalized data.

All the data used excludes self-funding by the candidates.

The data souce for the puropose of this Data project is FiveThirtyEight. Here is the list of variables that will be used for the project:

Value	Description
year	Election year
party	Political party
candidate	Candidate running in primary
endorsement_points	Weighted endorsements through June 30th of the year before the primary
percentage_endorsement_points	Percentage of total weighted endorsement points for the candidate's political party through June 30th of the year before the primary
money_raised	Money raised through June 30th of the year before the primary
percentage_of_money	Percentage of total money raised by the candidate's political party through June 30th of the year before the

	primary
primary_vote_percentage	Percentage of votes won in the primary
won_primary	Did the candidate win the primary?

Exploratory data analysis:

The explanatory variables in the data set are 'endorsement_points' and 'money_raised'. These are of type numerical. The response variable for the question we are trying to answer - Do number of endorsements increase the chances of winning the primaries? is 'primary_vote_percentage'.

Inference:

```
# Load the necessary packages to be used for this project
library(knitr)
library(ggplot2)
library(curl)
```

Load endorsements data - Upload the file to my github repository and use the curl package to retrieve the dataset in to R.

```
endorsements <-
read.csv(curl("https://raw.githubusercontent.com/isrini/SI_IS606/master/endorsements.csv"), header = TRUE)
kable(head(endorsements,10))
```

year	party	candidate	endorsement_points	percentage_endorsement_points	money_raised	percentage_of_money	primary_vote_percentage	won_primary
1980	Republican	George H.W. Bush	5	6.32911	1475332	16.83	23.81	No
1980	Republican	Lowell Weicker	0	0.00000	60000	0.68	0.00	No
1980	Republican	Phil Crane	0	0.00000	2500000	28.53	0.76	No
1980	Republican	John Connally	6	7.59494	2200000	25.10	0.64	No

1980	Republican	Bob Dole	0	0.00000	253286	2.89	0.06	No
1980	Republican	Ronald Reagan	58	73.41770	1400000	15.97	59.79	Yes
1980	Republican	Howard Baker	10	12.65820	643373	7.34	1.41	No
1980	Republican	John Anderson	0	0.00000	201745	2.30	12.19	No
1980	Republican	Larry Pressler	0	0.00000	0	0.00	0.00	No
1980	Republican	Harold Stassen	0	0.00000	30000	0.34	0.20	No

```
# correlation of endorsements to the vote percentage
cor(endorsements$percentage_endorsement_points,endorsements$primary_vote_percentage)

## [1] 0.7463713

# correlation of money raised to the vote percentage
cor(endorsements$percentage_of_money,endorsements$primary_vote_percentage)

## [1] 0.6055878
```

The endorsements have stronger correlation to the primary vote share compared to the money raised.

Let's also use the lm function in R to fit the linear model (a.k.a. regression line).

```
# linear model using endorsements
m1 <- lm(primary_vote_percentage ~ percentage_endorsement_points, data = endorsements)

# linear model using money raised
m2 <- lm(primary_vote_percentage ~ percentage_of_money, data = endorsements)
```

The first argument in the function lm is a formula that takes the form y ~ x. Here it can be read that we want to make a linear model of percentage endorsement points for m1 and percentage of money for m2 as a function of primary vote percentage received by each cnadidate. The second argument specifies that R should look in the endorsements data frame to find the respective variables.

The output of lm is an object that contains all of the information we need about the linear model that was just fit. We can access this information using the summary function.

```
summary(m1)
```

```
##
## Call:
## lm(formula = primary_vote_percentage ~ percentage_endorsement_points,
##     data = endorsements)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.611  -5.301  -3.437   2.063  48.273
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    3.63749    1.32510   2.745   0.0071 **
## percentage_endorsement_points  0.71048    0.06124  11.601   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.24 on 107 degrees of freedom
## Multiple R-squared:  0.5571, Adjusted R-squared:  0.5529
## F-statistic: 134.6 on 1 and 107 DF,  p-value: < 2.2e-16
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = primary_vote_percentage ~ percentage_of_money, data =
endorsements)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.902  -7.025  -2.801   1.387  49.157
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.75252    1.73531   1.586    0.116
## percentage_of_money  0.73167    0.09295   7.872 3.04e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.63 on 107 degrees of freedom
```

```
## Multiple R-squared:  0.3667, Adjusted R-squared:  0.3608
## F-statistic: 61.97 on 1 and 107 DF,  p-value: 3.035e-12
```

In the "Coefficients" table above; the first column displays the linear model's y-intercept and the coefficient of percentage_endorsement_points for m1 and percentage_of_money for m2. With this table, we can write down the least squares regression line for the linear model:

for m1:

$$\hat{y} = 3.6374 + 0.7104 * \text{percentage}_\text{e}\text{ndorsement}_\text{p}\text{oints}$$

for m2:

$$\hat{y} = 2.7525 + 0.7316 * \text{percentage}_\text{o}\text{f}_\text{m}\text{oney}$$

The R^2 value represents the proportion of variability in the response variable that is explained by the explanatory variable. For m1, 55.7% of the variability in vote percentage is explained by percentage of endordement points received. And for m2, 36.67% variablility is explained by percentage of money raised.

We see a positive relationship, for every 1 percent of endorsement points we would see an increase of 0.7104 percentage of primary vote. Similarly, for every 1 percent point of money raised we would see an increase of 0.7316 percentage of primary vote.

```
# To assess whether the linear model is reliable, we need to check for (1)
linearity,
# (2) nearly normal residuals, and (3) constant variability.

plot(m1$residuals ~ endorsements$percentage_endorsement_points)
abline(h = 0, lty = 3)  # adds a horizontal dashed line at y = 0

plot(m2$residuals ~ endorsements$percentage_of_money)
abline(h = 0, lty = 3)  # adds a horizontal dashed line at y = 0


# The function abline plots a line based on its slope and intercept. Here, we
used a
# shortcut by providing the model m1, which contains both parameter
estimates.

plot(endorsements$primary_vote_percentage ~
endorsements$percentage_endorsement_points)
abline(m1)

plot(endorsements$primary_vote_percentage ~ endorsements$percentage_of_money)
abline(m2)
```

The graphs above for the two linear models show that there is a linearity and the residual values are near normal showing constant variability.

The results indicate that the endorsements have higher correlation to the primary percentage votes gained compared to the money raised.

Conclusion:

There is evidence and correlation between endorsements and winning the preseidential primaries.

There are also some exceptions where neither the maximum number of endrsements or the amount of money raised in a given election primary year did not result in the wins. Hillary Cclinton in 2008 is a good example.

The primary election process is different than the general elections, primaries are a series of elections across each state to elect the nominee frome each of the two major parties. The candidates accrue points based on the states they carry. It is part of a long campaign process that usually takes many teists and turns.

Besides the chosen two variables there are many other variables like favourability, likely voters etc. that could impact the percentage votes gained. There needs to be firther study based on the data around other variables to arrive at a better conclusion as to how other variables could have played more significant part or not.

References:

https://github.com/fivethirtyeight/data/blob/master/endorsements-june-30/endorsements-june-30.csv