
ISyE 2028 – Basic Statistical Methods - Fall 2015

Bonus Project: “Big” Data Analytics

Final Report

Abstract:

The presidential election is one of the most anticipated events that occur every four years. Throughout the past few elections, candidates from both political parties have campaigned and debated extensively up until the final election. The current election has proven to be even more interesting with varying levels of popularity after every presidential debate. Polls are conducted before, during, and after each party debate to see how well-received each candidate is by the general public. It will be interesting to see how these numbers fluctuate throughout these past few months especially in determining who will become the frontrunners for each political party.

Methodology:

- Data retrieved from news outlets including ABC, CNN, FOX News, and MSNBC as well as compiled data from Polling Report and Real Clear Politics for the past 3 months with an average sample size of 412.
- 5-point numerical summaries for each of the 5 main presidential candidates.
- 95% confidence intervals based on recent aggregate polling as reported by Real Clear Politics.
- Linear regression models to demonstrate the popularity of each candidate within the polls.

Data Analysis:

Based on the data collected from various sources, it was clear to see how each candidate stood amongst their colleagues within the polls. Using a 5-point numerical summary, it was easily distinguishable as to which candidate stood a strong chance in obtaining the popular vote within their respective political party.

Min.	17.00
1 st Quartile	23.25
Median	25.00
Mean	25.26
3 rd Quartile	27.00
Max.	33.00
Var.	13.17212
Stand. Dev.	3.629341

Min.	5.00
1 st Quartile	12.25
Median	18.50
Mean	17.21
3 rd Quartile	22.00
Max.	29.00
Var.	43.73826
Stand. Dev.	6.613491

Min.	4.00
1 st Quartile	7.00
Median	8.00
Mean	8.447
3 rd Quartile	9.750
Max.	15.00
Var.	8.037696
Stand. Dev.	2.835083

Min.	18.00
1 st Quartile	25.00
Median	31.00
Mean	29.36
3 rd Quartile	33.00
Max.	38.00
Var.	26.05195
Stand. Dev.	5.104111

Min.	33.00
1 st Quartile	45.75
Median	55.00
Mean	52.32
3 rd Quartile	56.75
Max.	64.00
Var.	56.22727
Stand. Dev.	7.498485

Based on these numerical summaries, it can be concluded that Donald Trump and Hilary Clinton have the greatest popularity within their respective parties considering that the mean percent of voters for either candidate is greater than the mean of their competitors. In addition, based on these summaries, it can be seen that candidates such as Jeb Bush and Donald Trump with a relatively low standard deviation will remain where they are in the polls whereas those with a greater value may or may not move in ranks. Using this information, it can be concluded that certain candidates have a strong standing within their political party.

In order to analyze how each political debate influenced a candidate's poll results, data was collected from poll data before and after each political debate. From the data, the following table of mean values is constructed to demonstrate the difference public opinion before and after the candidates have debated:

Republican Candidates	Debate #1: Aug. 6th		Debate #2: Sept. 16th		Debate #3: Oct. 28th		Debate #4: Nov. 10 th	
	Before	After	Before	After	Before	After	Before	After
Donald Trump	22.86	27.83	30.67	24.22	26.33	24.50	26.00	28.00
Ben Carson	6.43	15.67	20.67	17.44	22.17	24.75	23.00	20.00
Jeb Bush	12.29	8.50	7.67	7.56	7.17	8.75	4.00	5.67

Democratic Candidates	Debate #1 Oct. 13th		Debate #2: Nov. 14th	
	Before	After	Before	After
Bernie Sanders	26.33	31.11	33.00	32.00
Hilary Clinton	46.44	56.78	52.00	56.67

As seen through these averages, the public opinion tends to increase after each political debate particularly for Hilary Clinton who is the current front-runner for the Democratic Party nomination. Depending on their performance in the debate, the public tends to vote favorably for the candidate. However, in some cases such as that of the Republican candidates, the poll results have looked unfavorably on the candidate's popularity after the debate.

Recent aggregate polling demonstrates the average on which each candidate obtain a certain percent of voters in a poll based on three common polls: Fox News, ABC/Washington Post, and Public Policy Polling. In order to test the confidence of these averages, 95% confidence intervals were conducted based on the data collected in the past 3-4 months. The following averages are listed for each presidential candidate:

Republican Candidates	Donald Trump	28.7%
	Ben Carson	19.7%
	Jeb Bush	5.3%
Democratic Candidates	Bernie Sanders	30.2%
	Hilary Clinton	55.8%

Using the numerical summaries, the following 95% confidence intervals were calculated:

$$P\{L \leq \mu \leq U\} = 1 - \alpha$$

$$P\left\{\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

$$CI: \left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right]$$

Given that $Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$ and $n = 37$ for the Republican candidates while $n = 21$ for the Democratic candidates, the confidence interval for the first Republican candidate, Donald Trump is calculated as so:

$$25.26 - 1.96 \frac{3.63}{\sqrt{37}} \leq \mu \leq 25.26 + 1.96 \frac{3.63}{\sqrt{37}}$$

$$24.0935 \leq \mu \leq 26.430$$

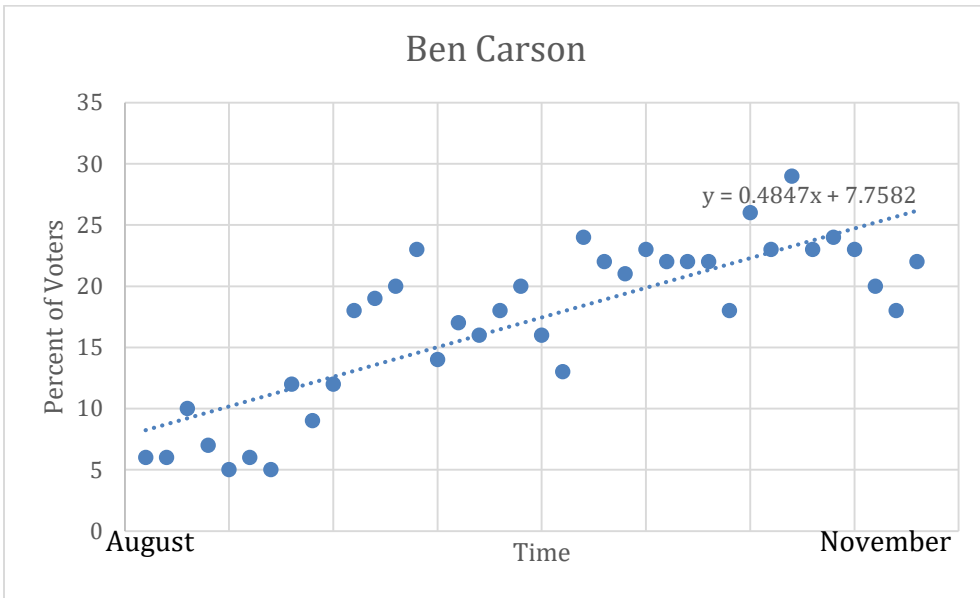
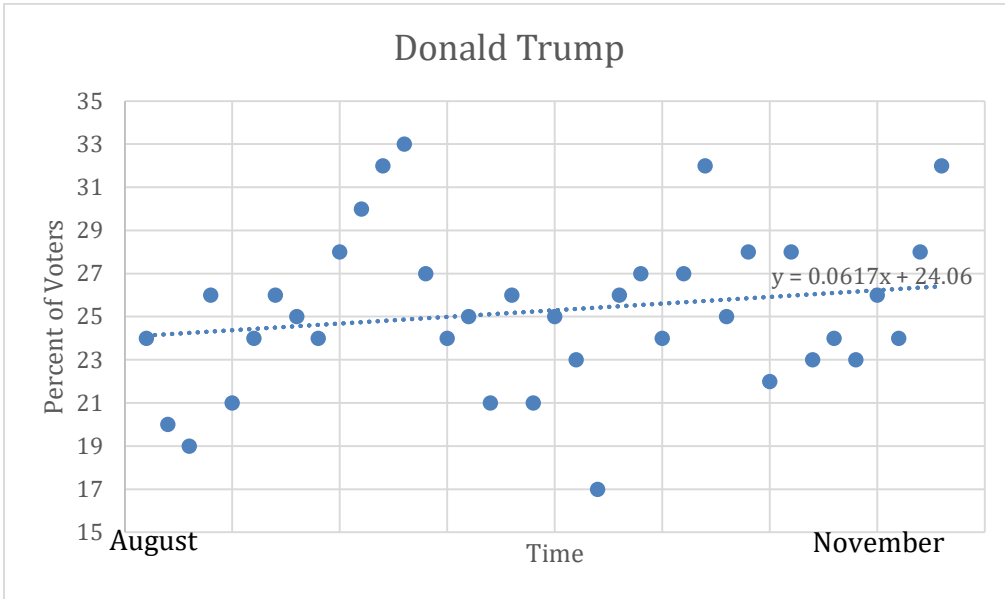
Given this interval, the aggregate poll value does not fall within this range indicating that the average does not adequately represent the percentage of voters that will vote for Donald Trump. Rather, if another poll is taken of a random sample, then it's 95% confident that the percentage of voters for Donald Trump will fall within this range. Similarly, the confidence intervals for the other candidates were also computed.

Republican Candidates	Donald Trump	$24.0935 \leq \mu \leq 26.430$
	Ben Carson	$15.03673 \leq \mu \leq 19.38433$
	Jeb Bush	$7.5155 \leq \mu \leq 9.379237$
Democratic Candidates	Bernie Sanders	$27.10060 \leq \mu \leq 31.62667$
	Hilary Clinton	$48.99354 \leq \mu \leq 55.64282$

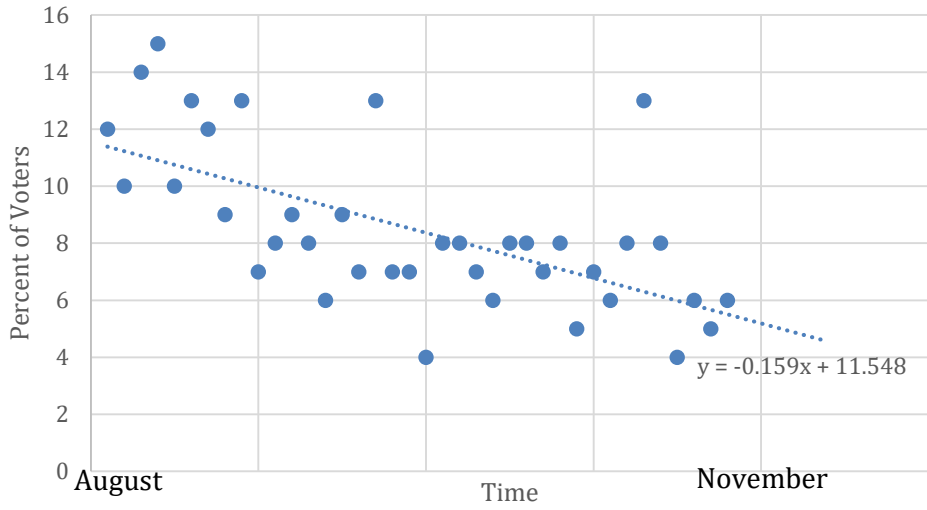
*Values computed using R

In addition to determining the confidence intervals for poll results for each candidate, linear regression models help present a visual depiction as to the candidates' position within political

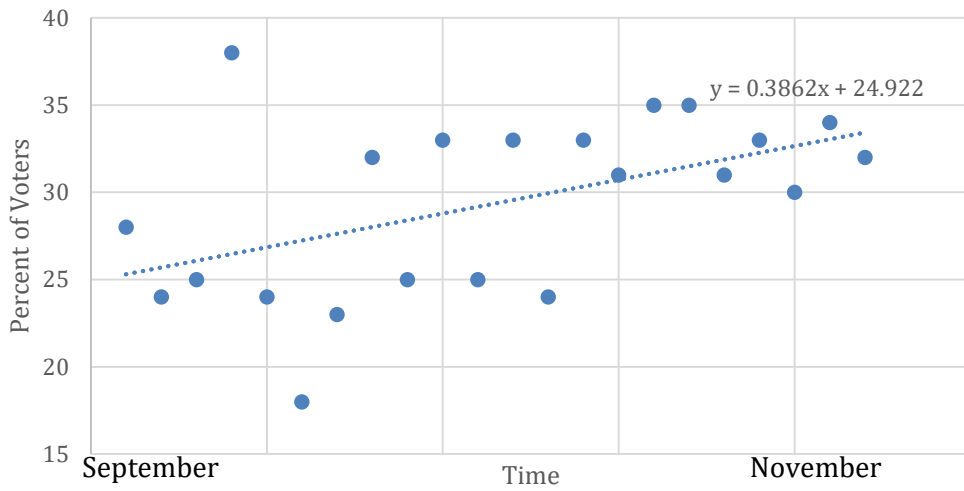
polls. Based on a sequential analysis of various polls taken nationwide, each model was constructed to denote each candidate's progress on the campaign trail.

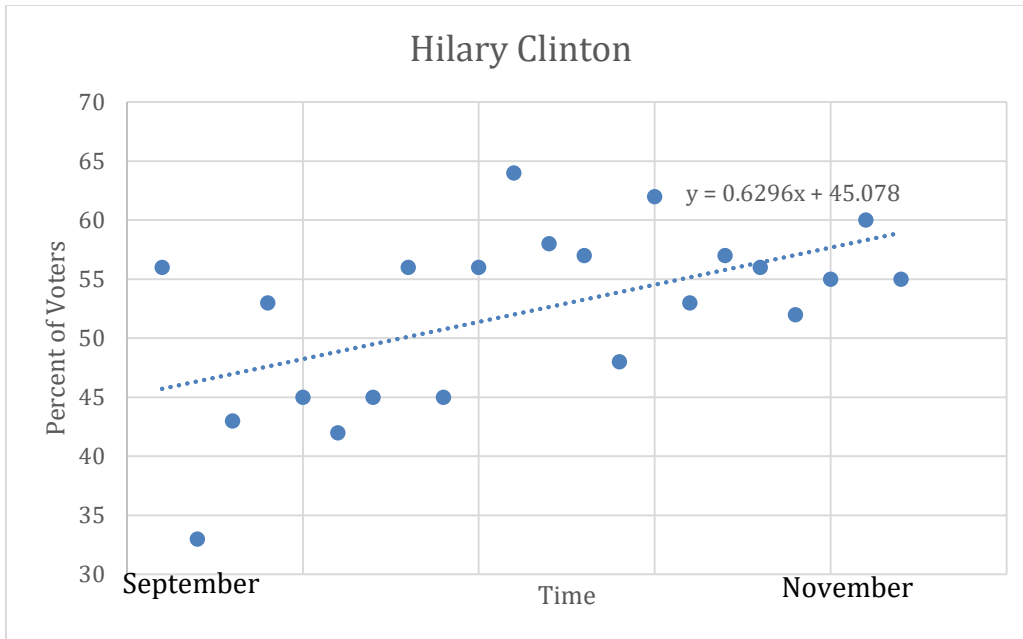


Jeb Bush



Bernie Sanders





Based on these linear regression models, certain candidates follow a more consistent trend than others in terms of poll results.

Conclusion:

From the data retrieved, it can be concluded that the front-runners for the Republican and Democratic parties are Donald Trump and Hilary Clinton, respectively. From the numerical summaries conducted, both candidates polled on average higher than their contenders. In addition, their confidence intervals indicate that these candidates will consistently poll at these higher averages thereby indicating their general popularity within the polls. Considering the averages from before and after each political debate, the candidates that have a consistent trend in the polls also do considerably well after the debate. For example, Donald Trump has generated a rather consistent standing within the polls which is reflected by the positive linear trend of the linear regression model as well as a general increase in the percentage of voters before each Republican debate as referenced in Table 6.

In contrast, the Democratic candidates have had a consistent increase in their poll results than that of the Republican candidates. In comparing the linear regression models, Ben Carson has had the highest increase in the polls whereas Jeb Bush has had the largest decrease. This may be due to the increasing popularity Ben Carson has garnered after each political debate. There is an overall increase in percentage of voters for Ben Carson compared to the decrease in voters for Jeb Bush. In addition, it is surprising to note that Hilary Clinton has had a general increase in the poll results as well as Bernie Sanders considering that both are strong contenders for the Democratic nomination.

In conclusion, the poll results demonstrate that there are definite rankings for each of the candidates. However, public opinion is always fluctuating and in order to develop a more concrete ranking of each candidate, polls with larger sample sizes are needed in order to represent larger portions of the population. In addition, more frequent polls are needed to ensure that poll averages significantly represent the percentage of voters for each candidate.

References:

- "2016 Democratic Presidential Nomination." *RealClearPolitics - Election 2016*. RealClearPolitics.com, 2015. Web. 29 Nov. 2015. <http://www.realclearpolitics.com/epolls/2016/president/us/2016_democratic_presidential_nomination-3824.html>.
- "2016 Republican Presidential Nomination." *RealClearPolitics - Election 2016*. RealClearPolitics.com, 2015. Web. 29 Nov. 2015. <http://www.realclearpolitics.com/epolls/2016/president/us/2016_republican_presidential_nomination-3823.html>.
- "White House 2016: Democratic Nomination." *PollingReport.com*. POLLING REPORT, INC., 2015. Web. 29 Nov. 2015. <<http://www.pollingreport.com/wh16dem.htm>>.
- "White House 2016: Republican Nomination." *PollingReport.com*. POLLING REPORT, INC., 2015. Web. 29 Nov. 2015. <<http://www.pollingreport.com/wh16rep.htm>>.