Evaluating the Connection Between Internet Coverage and Polling Accuracy

California Propositions 2005-2010

Erika Oblea
December 12, 2011
Statistics 157
Professor Aldous

Introduction:

Polls are often taken prior to the election to predict the outcome of the election. The statistics behind polling is fairly established. People are randomly sampled and asked through sometimes carefully crafted unbiased questions how they would vote on Election Day. However, while the mathematical theory behind polling is nearly fool-proof, polls do not always get it right. One of the biggest factors in skewing polling accuracy is the problem of nonattitudes (Asher 32). This problem essentially exists because lack of voter knowledge on the issue being polled will often lead voters to haphazardly choose a response simply to appear informed on the issue. Herbert Asher claims that when voters being polled are confronted with a question that they know little about, very few voters will want to admit they know nothing on the issue. In this way, whatever mathematical certainty may be used to justify polling results, polls may simply just be a reflection of haphazard voter opinion and not a true reflection of how voters will actually vote on Election Day.

This project seeks to understand how this problem of nonattitude manifests itself with California propositions. Propositions are proposed laws that are submitted directly to the electorate to be voted into law or not. California is one of the few states in the United States that allows voters to vote directly on state legislation. Because often times these propositions entail confusing and obscure measures, information about these propositions may be crucial in solidifying voter knowledge on these propositions and consequently, polling accuracy on these propositions. According to Marcus Prior, the most influential factor these days in increasing voter knowledge on political issues like propositions is the Internet as the plethora of media outlets and blogs on the Internet has increased voter exposure to political information (Prior 577-8). Stephen P. Nicholson, in fact, performed his own statistical research confirming that media coverage, especially in the form of newspaper coverage, often increases voter awareness of propositions in California by nearly 16 percent. His research obtained this estimate by looking at how many front page references a proposition had in the *Los Angeles Times* prior to the election and relating that result with the percentage of people who knew about the proposition (Nicholson 407-8).

Using his research as a basis, this project seeks to understand how voter knowledge on California propositions can influence the accuracy of polls in predicting the outcome of these propositions. Since voter knowledge on a proposition is often a function of how much Internet coverage is available on the proposition, this project seeks to perform an exploratory data analysis to understand the relationship between Internet coverage on California propositions and polling accuracy on those propositions. Does having more information on propositions on the Internet allow polling to better predict election results on those propositions? If so, does more formal sources of media, like newspapers, make for more accurate polls? Or is it the more informal sources of media like blogs that increase polling accuracy? Do certain types of propositions elicit more media attention and thus better polling accuracy?

How the Data Was Collected:

Polling information on California propositions was retrieved from The Field Poll's website. The Field Poll is an independent and non-partisan media-sponsored public opinion news service that often focuses on the political issues surrounding California. Since 1994 until 2010, they have polled voter opinion on certain California propositions and then compared how their polling matched with the actual election outcomes on the proposition. From this data, I collected information on the year the proposition was voted, the proposition's name, the final Field Poll prediction on the proposition's outcome, and the actual outcome of the proposition in the election. The final Field Poll prediction is a series of percentages that indicated how people responded to whether they would vote for the proposition, either Yes, No, or Undecided. These predictions were obtained from a survey that was taken a week before the election.

The actual outcome of the election in the proposition is a pair of percentages, indicating Yes or No, on how people actually voted on the propositions on Election Day. The Field Poll considered its polling to be consistent for a specific proposition if the general direction of public opinion towards the propositions as measured by their polls matched the actual outcome of that proposition in the election. Although the Field Poll did not conduct polls for all of the propositions being voted on in a particular year, it claims that it did conduct polls on the most salient propositions that year.

From this data, I calculated the polling margins of the propositions by subtracting the percentage of people who responded "Yes" to the survey from the percentage of people who responded "No". I decided to eliminate the percentage of "Undecided" from this calculation since Field Poll does not utilize this information to determine how consistent their polling was in predicting election outcome. To calculate the election margins of the propositions, I subtracted the percentage of people who voted "Yes" from the percentage of people who voted "No" on the proposition on Election Day. Polling accuracy was measured by taking the absolute value of the difference between election margins and polling margins.

In order to determine Internet presence, I looked at how many blogs and newspaper records online referenced an individual proposition prior to when Field Poll conducted its survey. Using Google Blog Search, I searched for "California Proposition [Proposition Number]" for each proposition. The number of search hits composed the blog count for each proposition. To collect newspaper counts, I looked at four different newspaper websites, two national newspapers (*New York Times* and *USA Today)* and two more California-based newspapers (*San Francisco Chronicle* and *Los Angeles Times*) in order to get a more holistic sense of online newspaper coverage. On each newspaper's online website, I searched for "California Proposition [Proposition Number]" for each proposition. The sum total for the total number of hits for each newspaper search composed the newspaper count. I limited all of my newspaper and blog searches to the dates July 1 to October 28 of the year the proposition was voted. This timeframe represents the four months prior to the week of the election when Internet presence could have influenced public opinion before Field Poll conducted their survey.

From this data, I decided to simply look at the years 2005 to 2010. Although the Internet had become available to the public as early as 1994, Internet usage did not become quite as popular until much later. According to the World Bank, only 4.9% of the U.S. population was using the Internet in 1994. It was not until 2005 when much more than a majority of the U.S. was using the Internet (69.6%) (*Internet Users (Per 100 People)*). In this way, it would be uninformative to look at how Internet presence affected polling accuracy as early as 1994. Although 2004 was arguably the year that blogs became more mainstream (Rosenberg), I decided to utilize 2005 as a starting point because using this year may better ensure that blogs were a more established forum for political discussion.

All in all, my data consists of information for 43 propositions dating from 2005 to 2010. This data used in this project are the polling margins, the election margins, the polling accuracies, the blog counts, and the newspaper counts for each of these propositions.

Observations:

Consistent versus Inconsistent Polling

The first question is whether there is a fundamental difference in the amount of Internet attention between the propositions the polls were able to accurately predict and the propositions the polls were not able to accurately predict the outcome. To answer the question, I performed a Chi-Squared Test of Homogeneity. I am assuming that the collection of newspaper and blog counts consist of independent samples from the set of propositions with consistent polling and the set with inconsistent polling.

The null hypothesis is that there is no difference in the Internet presence of the propositions between the propositions the polls accurately predicted versus the ones the polls did not accurately predict. The alternative hypothesis is that there is a difference. The Chi-Squared statistic was calculated to be 31.5185 and the corresponding p-value was 1.975e-08. Therefore, at the 1% significance level, we can reject the null and claim there is most likely a difference in Internet presence between the two types of propositions.

More importantly, when we take a look at contributions to the chi-square statistic cell-by-cell, we see that the presence of online newspaper coverage seems to have the values that contribute most to the chi-squared statistic. This suggests that the presence of online newspaper coverage may be a defining factor differentiating the propositions the polls could accurately predict and the ones when they could not. The tables below demonstrate that propositions with consistent polling tend to be associated with more newspaper coverage than propositions with inconsistent polling.

Observed vs. Expected Values

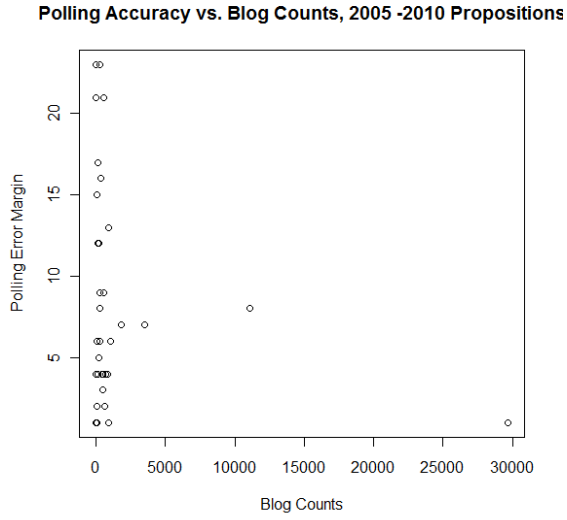
	Consistent	Inconsistent
Newspapers	235 184.6528	88 138.3472
Blogs	56809 56859.3472	42651 42600.6528

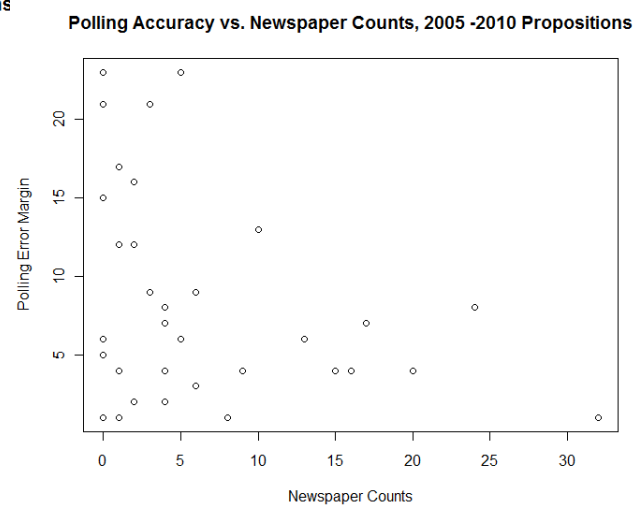
Contributions to chi-squared statistic

	Consistent	Inconsistent
Newspapers	13.72759388	18.32230200
Blogs	0.04458086	0.05950235

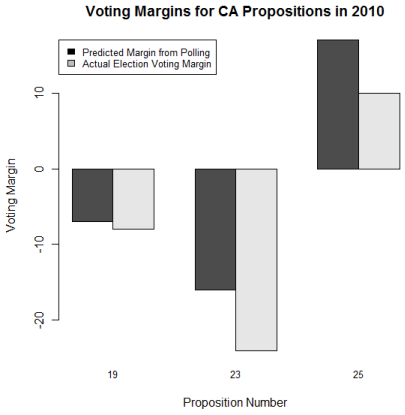
Consistent Polling

The last section established that there does seem to exist a difference between consistent polling and inconsistent polling with regards to the Internet presence of the propositions the polls were predicting. This section will now evaluate the relationship between polling accuracy and Internet presence of the propositions with just the propositions the polls were able to accurately predict. The following analyses involve a subset of the dataset. It mostly looks at the data where the polls accurately predicted the outcome of the proposition in the election.

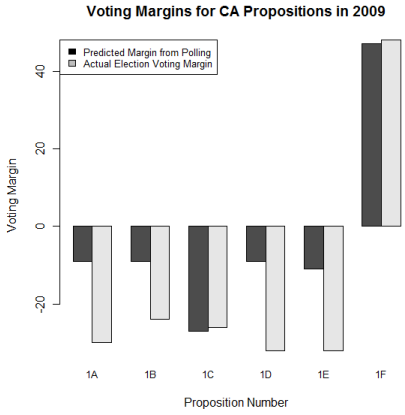




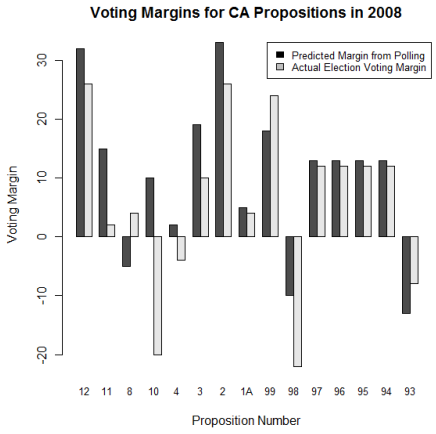
Looking at the scatterplots above, it does not seem there is a purely linear or even a discernable relationship between the amount of blogs or newspapers and the corresponding polling margin of error. It may be inappropriate to fit a linear model to this data. It then may be important to analyze the relationship not purely within the whole five year time frame but rather on a year-by-year basis.



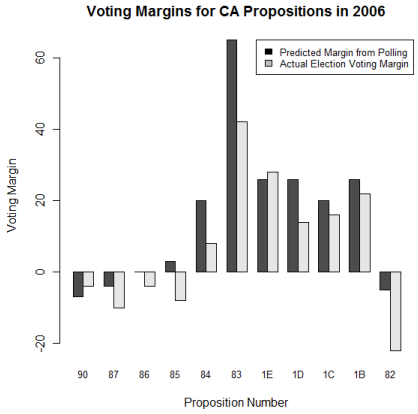
mean diff: 6.375



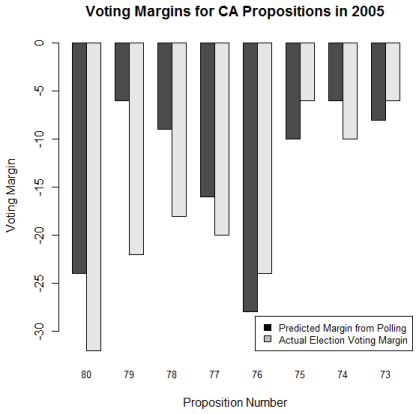
mean diff.: 8.909



mean diff: 7.20



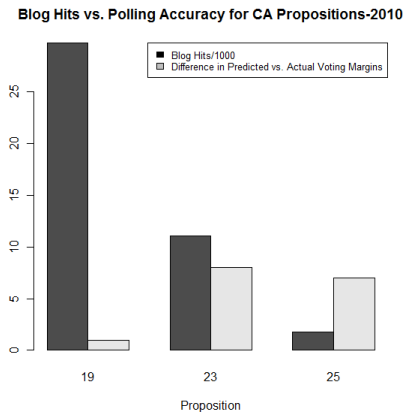
mean diff: 13.667

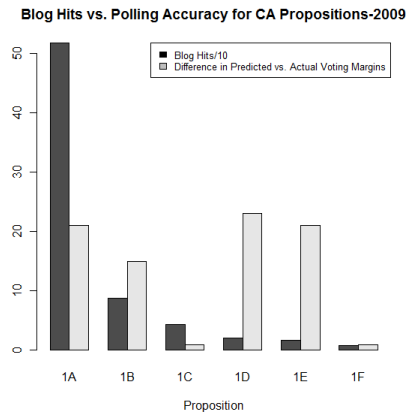


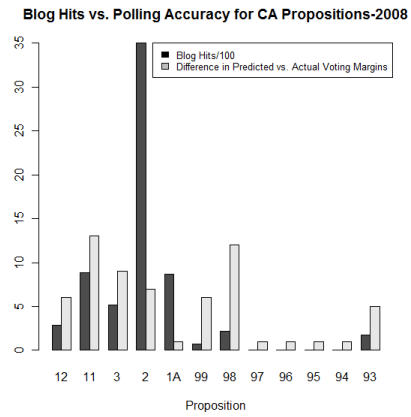
mean diff: 5.3333

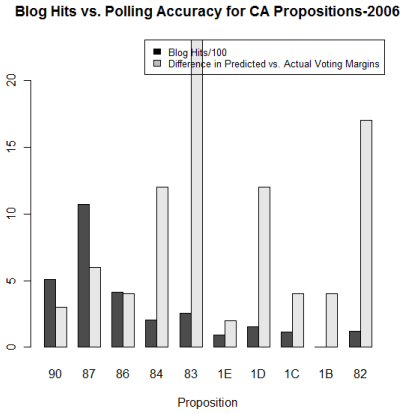
The barcharts demonstrate that the difference between predicted and actual voting margins for the propositions varies from year to year. The years with the largest mean difference between predicted and actual voting margins for propositions with accurate polling occur in 2006, 2008, and 2009. Not surprisingly, 2006 and 2008 are also the years when polling was not able to accurately predict the outcome for some propositions (Proposition 8,10,4 in 2008 and Proposition 85 in 2006). 2005 and 2010 both have the propositions with the lowest margin of polling errors. Although there are some propositions in 2005, such as 79 and 78, that have very discrepant predicted and actual voting margins, polling was able to very closely predict most of the other voting margins for the propositions. Furthermore, during 2005 and 2010, the polls were still able to accurately predict all the outcomes of the propositions.

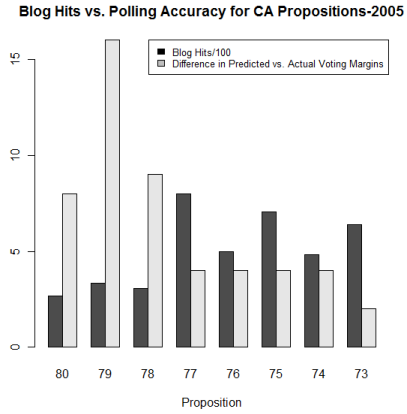
In this way, these graphs demonstrate that each year from 2005 to 2010 represents a different political climate under which these propositions were voted. It is for this reason it may be more insightful to evaluate the effect of Internet discussion on these propositions on a year-by-year basis rather than on the whole period ranging from 2005 to 2010.





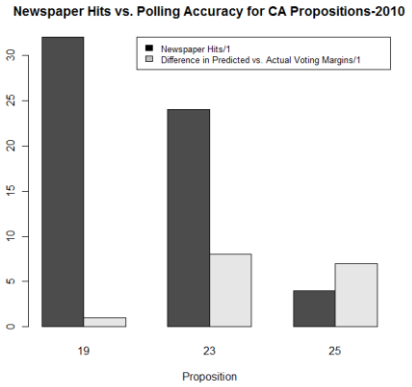


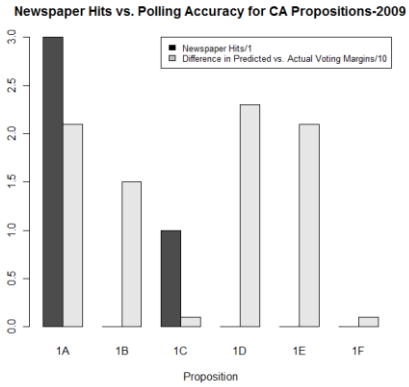


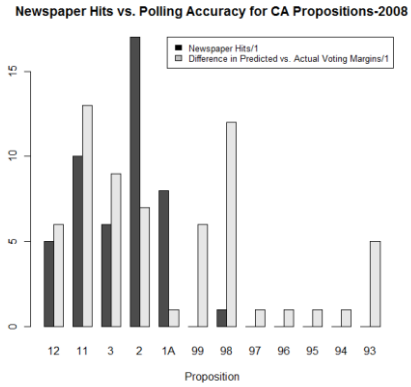


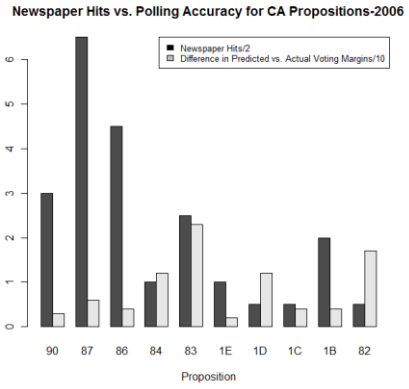
In 2005 and 2010, the years with the lowest average difference between predicted and actual voting margins, there is a much clearer relationship between the frequency of blog discussion and the accuracy of the poll in predicting the outcome of the proposition. For instance, in 2010 and 2005, the propositions with the most amount of blog discussion prior to the election (Proposition 19 in 2010 and 77 in 2005) also had the polls with one of the most accurate predictions. On the other hand, the propositions with the least amount of blog discussion in 2005 and 2010 (Proposition 79 and 80 in 2005 and Proposition 25 in 2010) had some of the least accurate polls.

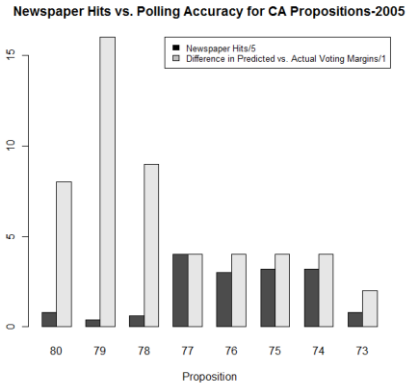
In 2006, 2008, and 2009 this relationship is not as clear. For example, in 2009, Proposition 1D and 1E had the least amount of blog discussion at the time but still had nearly the same polling accuracy as Proposition 1A which had the most blog discussion that year. The same phenomenon could be said for the propositions in 2006 and 2008 as a larger amount of blog discussion does not seem to be associated with more accurate polling. In 2008, Proposition 12 had much less blog discussion than Proposition 2 of that year but both propositions had nearly the same polling accuracy. In 2006, Proposition 87 had much more blog discussion than Proposition 1C that year but again both propositions had nearly the same polling accuracy.



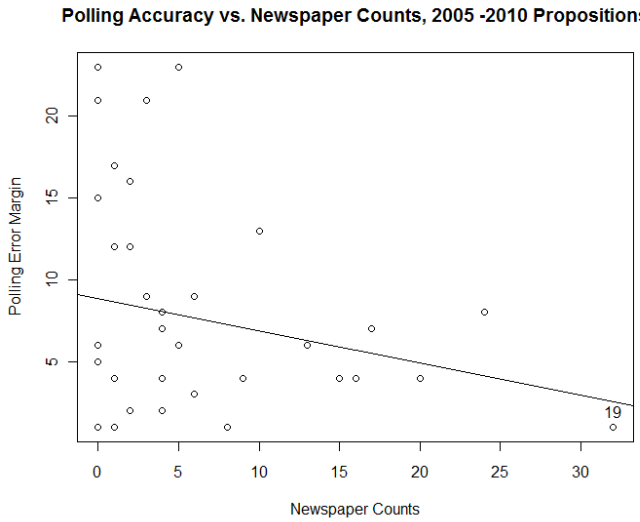


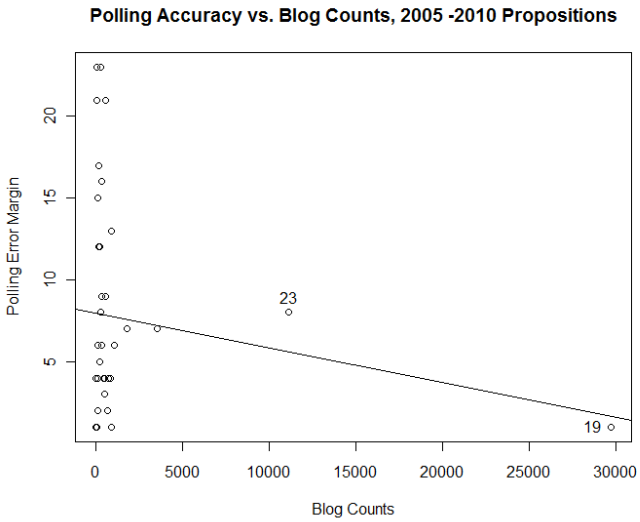






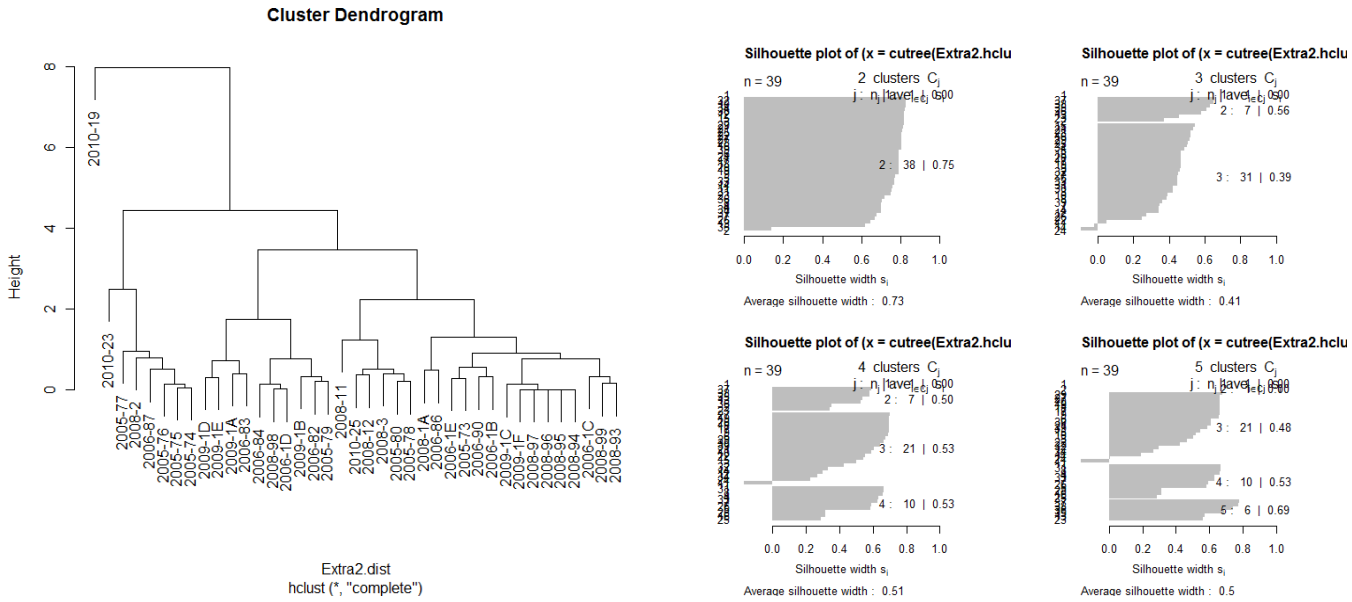
In 2005 and 2010, there does seem to be a relationship between the amount of online newspaper coverage and polling accuracy. For instance, the propositions with the most online newspaper coverage that year prior to the election (Proposition 19 in 2010 and Proposition 77 in 2005) also had some of the most accurate polls that year. At the same time, the propositions with the least amount of online newspaper coverage that year also had some of the most inaccurate polls that year (Proposition 25 in 2010 and Proposition 79 in 2005). This relationship, however, is not so clear when it comes to the years 2006, 2008, and 2009. For example, in 2009, Proposition 1A had the most online newspaper discussion that year but its polling accuracy is nearly the same as that for Proposition 1D and 1E of that year that had no online newspaper discussion. This pattern can also be verified in 2006 and 2008. In 2008, Proposition 2 had the most online newspaper coverage that year but it had nearly the same polling accuracy as Proposition 99 that year that had no online newspaper coverage. In 2006, Proposition 87 had the most online newspaper coverage that year but had nearly the same polling accuracy as Proposition 1C that had some of the least newspaper coverage of that year.





To understand whether the number of newspapers or blogs is a better predictor of polling accuracy, it might be relevant to compare the linear model for polling accuracy vs. the number of blogs and the linear model for polling accuracy vs. number of newspapers. The following information, however, should be taken very lightly as the plots do not seem to follow any linear pattern. When fitting a linear model to the data for polling accuracy vs. newspaper count, the linear relationship is negative (slope = -.1967) but very weak (R^2=.05182). When fitting a linear model to the data for polling accuracy vs. blog count, the linear relationship is also negative (slope = -.0002119) but also very weak (R^2 =0.02529). Furthermore, this linear model seems to be a lot weaker because it is driven a lot by two outliers in the scatterplot (Proposition 23 and Proposition 19). However, the R^2 statistic is larger for the linear model fitting polling accuracy vs. newspaper count. This suggests that the number of newspaper counts may have a slightly stronger association with polling accuracy, albeit a weak one. More specifically, the more online newspaper discussion there is, the more accurate the poll is in predicting the outcome of the proposition in the election.

In order to understand how this relationship holds between different kinds of propositions, I performed a cluster analysis. A cluster analysis essentially looks at several observations when there are several available measurements. In this case, the observations are the 43 propositions voted on between 2005 and 2010 and the measurements are the polling accuracy of these propositions as well as the number of blogs and newspaper articles mentioning the individual propositions before the election. Using these measurements, a cluster analysis attempts to determine whether the observations naturally group together in some predictable way. While there are many methods of clustering, I used the hierarchal agglomerative clustering method. This method begins by putting each observation into its own cluster. It then calculates the distances between all observations and pairs together the two observations that have the smallest distances to form a new cluster. The mean of this new cluster is then calculated and the distance between this mean and the other observations is calculated. The process then repeats itself until all observations are properly placed into clusters.



After applying hierarchal agglomerative clustering to the data, I produced the dendogram plot above which essentially summarizes how the hierarchal agglomerative clustering method suggests grouping the data. Looking at the dendrogram, it looks like the observations naturally fall into four different groups. This is confirmed by the silhouette plot on the right. The silhouette plot with one of the largest average silhouette width is the one where the observations are grouped into four clusters. The average silhouette width is .51, which by clustering standards suggests that a reasonable structure has been found. The scale for an average silhouette width is from 0 to 1, with 0 being no fit and 1 being a perfect fit. Now it is essential to understand what factors cause the observations to fall into four groups according to hierarchal agglomerative clustering.

Cluster	Frequency	Mean Polling Accuracy	Mean Blog Count	Mean Newspaper Count
1	1	1.000000	29700.0000	32.000000
2	7	5.285714	2593.7143	17.285714
3	21	4.238095	334.3810	3.190476
4	10	17.200000	193.1000	1.500000

The four groups seem to be divided based on the level of polling accuracy, blog count, and newspaper count. More specifically, group 1 seems to contain 1 observation with the best polling accuracy and highest blog count and newspaper count while group 4 seems to contain 10 observations with the worst polling accuracy and smallest blog and newspaper counts. Group 2 and 3 seem to contain moderate polling accuracy, blog count, and newspaper count. There are 7 propositions in Group 2 and 21 propositions in Group 3.

[[1]] Marijuana Legalization

[[2]] Suspension of AB32	Farm Animals
Oil Producers Tax	Cigarette Tax
Redistricting	Spending / School Funding Limits
Union Dues	Teacher Tenure

[[3]] Majority Vote – State Budgets	Lottery Modernization
Elected Officials' Salaries	Veterans' Bond
High-speed Rail Bond	Eminent Domain
Indian Gaming	Indian Gaming
Indian Gaming	Indian Gaming
Term Limits Reform	Eminent Domain
Disaster Preparedness/Flood Bond	Housing Bond
Transportation Bond	Electricity Regulation
Parental Notification of Teen Abortion	Budget, budget, salaries, term limits
Bonds, bonds, eminent domain	Indian gaming
Electricity regulation	Parental notification

[[4]] State Spending Limit; Rainy Day Fund	Education Funding
Children's Services Funding	Mental Health Funding
Redistricting	Children's Hospital Bond
Eminent Domain	Water Quality/Parks Bond
Sex Offenders	Education Facilities Bond
Pre-School Education	Rx Drug Discounts (II)
Rx Drug Discounts (I)	

When we look at how the hierarchal agglomerative clustering method divided the propositions into four groups and look at how the four groups are divided by the proposition's topic, nothing really stands out. Each group does not seem to contain one kind of proposition topic over another. For instance, some education-related propositions are found in Group 2 (Spending/School Funding Limits, Teacher Tenure) and in Group 4 (Education Funding) while more legislative reform propositions are found in Group 2 (Redistricting), Group 3 (Elected Officials' Salaries, Majority Vote-State Budgets, Term Limits Reform), and in Group 4 (Redistricting, State Spending Limit; Rainy Day Fund).

However, if we look at the names of the propositions themselves, we do seem to see some kind of coherent division of groups. The computer output below looks at which proposition numbers are placed into each group by hierarchal agglomerative clustering. The format is "Year – Proposition Number".

[[1]] "2010-19"

[[2]], "2010-23" "2008-2" "2006-87" "2006-86" "2005-77" "2005-76" "2005-75" "2005-74"

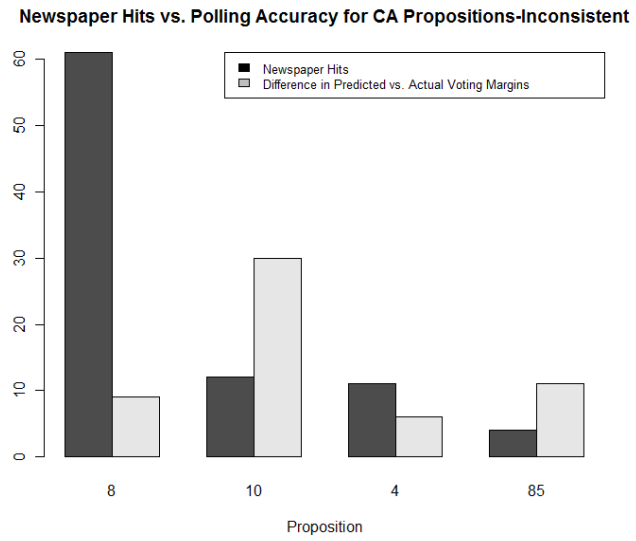
[[3]] "2010-25" "2009-1C" "2009-1F" "2008-12" "2008-1A" "2008-99" "2008-97" "2008-96" "2008-95" "2008-94" "2008-93" "2006-90" "2006-1E" "2006-1C" "2006-1B" "2005-80" "2005-73"

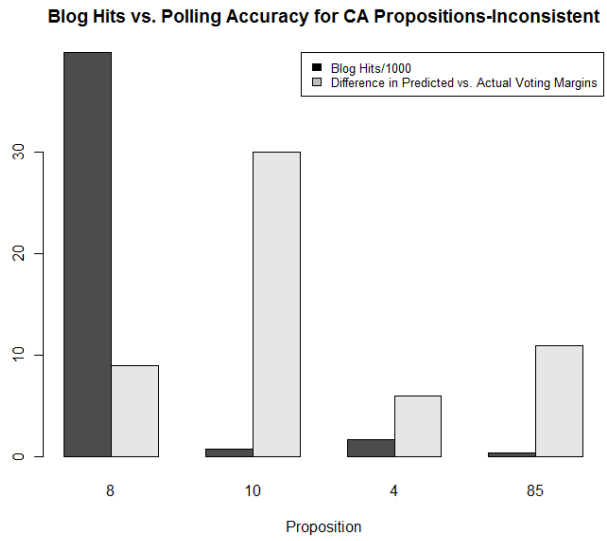
[[4]] "2009-1A" "2009-1B" "2009-1D" "2009-1E" "2008-11" "2008-3" "2008-98" "2006-84" "2006-83" "2006-1D" "2006-82" "2005-79" "2005-78"

If the Proposition number has a letter attached to the end, that means the proposition was a legislative-initiated proposition. This means that the proposition was first introduced in the legislature and was only put to a direct vote by the populace because it was not voted in the legislature. It seems that all of the legislative-initiated propositions fall into Groups 3 and Groups 4. In fact, nearly 50 percent of the members of Group 4 are legislative-initiated propositions. This suggests that the legislative initiated propositions tend to be the propositions that have polls that most inaccurately predict the proposition's outcome and which also have some of the least amount of blogs or newspaper content online.

What is also interesting to note from this cluster analysis is that Proposition 19 from the 2010 election is placed into a group of its own. Proposition 19 seems to form an extreme outlier as Proposition 19 seems to have the largest amount of blog counts and online newspaper content. Furthermore, the poll predicting the outcome of Proposition 19 is the most accurate. In fact, Proposition 19 also appeared as an outlier in the scatterplots plotting polling accuracy vs. blog counts and polling accuracy vs. newspaper counts. The fact that the relationship between Internet presence and polling accuracy is so strong for Proposition 19 may be a result of the proposition's topic. The issue of marijuana legalization was a particularly heated topic in the 2010 election ("Marijuana and Medical Marijuana"). It was only natural that there was much online debate on the issue and that people were heavily opinionated on the topic, thus improving the poll's accuracy in predicting the proposition's outcome.

Inconsistent Polling





These barcharts present an interesting relationship between the number of blogs and online newspaper discussion and polling accuracy for the propositions the polls did not accurately predict the outcome. Namely, the more Internet discussion there was, the less inaccurate the poll was. This is evident for Proposition 8 in 2008. On the other end, the less Internet discussion there was on the proposition, the more inaccurate the poll was in predicting the election for the proposition. This is most evident for Proposition 10 in 2008. Even though the polls did not accurately predict the election's outcome, the relationship between Internet discussion and polling accuracy is nearly the same as when the polls do accurately predict the election's outcome.

Conclusion:

In conclusion, the observations from this project illustrate that the relationship between polling accuracy and Internet coverage may not be completely clear cut. It is not obvious from the data that in all cases more Internet coverage on a particular proposition generated polls that better predicted the outcome of that proposition in the election.

On one hand, there does seem to be a difference in Internet coverage on propositions when it comes to determining whether or not a poll accurately predicted the outcome of that proposition. Consistent polls are more heavily associated with more Internet coverage, suggesting that Internet coverage may be a factor in helping polls better predict the outcomes of propositions in California.

When we take a look at the propositions the polls were able to accurately predict, we find that there is no clear relationship between polling accuracy and Internet coverage when looking at the entire timeframe from 2005 to 2010. However, when we look at the data on a year-by-year basis, we find that more Internet coverage was associated with more accurate polls for propositions in 2005 and 2010. In 2006, 2008, and 2009, this relationship is not so clear. Furthermore, we find that more formal sources like online newspaper coverage may be more influential in determining polling accuracy than more informal sources like blogs. A proposition's type also seems to influence the relationship between Internet coverage and polling accuracy. Namely, the legislative-initiated propositions seem to be associated more with less Internet coverage and less polling accuracy. This idea seems logical given that the public would be less invested in a proposition they did not initiate themselves and would thus have less interest in discussing it and would thus have less opinion on it.

However, what is most interesting is that Proposition 19 continually emerges from this data as a significant outlier in that it was a proposition with an inordinate amount of Internet coverage and was also a proposition the polls were able to very accurately predict. This information makes sense given that Proposition 19 represented the first time in nearly four decades California voters were deciding whether to legalize marijuana, a political issue that taps into a lot of deep-seated moral viewpoints on drug usage in this country. When it comes to looking at the propositions the polls could not accurately predict, we find a similar relationship. Namely, the more Internet coverage there was on that poll, whether newspaper or blog, the less inaccurate that poll was.

While this research project produced data that illustrated some interesting caveats to the relationship between Internet coverage and polling accuracy for California propositions between 2005 and 2010, this project could still benefit from some improvements. To start, the information was taken solely from Field Poll which only polled a select number of salient propositions. Field Poll is only one of many polling organizations in California, and future research could possibly benefit from data that incorporated the polling accuracies for propositions from other polling firms. This research could also be further strengthened if information on polling accuracies and Internet coverage could be gathered for every single proposition that was voted by Californians between 2005 and 2010. In addition, the method used to gather blog counts and newspaper counts for each proposition was not fool-proof. Although much care was taken to ensure that the searches only displayed relevant results, the counts used in this project may still not be completely accurate representations of the Internet coverage for the propositions at the time. A future project could better determine how to produce more accurate counts.

All in all, however, this research project demonstrated that Internet coverage to a certain extent may influence how accurately polls can predict the outcome of a proposition in the election for California. This research project validates the notion that the more informed voters are on a proposition in California, the more likely public opinion polls can better reflect voter behavior on Election Day.

Works Cited

Asher, Herbert. *Polling and the Public*. Washiington, D.C: CQ Press, 2007. Print.

Internet Users (Per 100 People). World Bank, n.d. Web. 5 Dec. 2011. <http://data.worldbank.org/indicator/IT.NET.USER.P2?page=3>.

"Marijuana and Medical Marijuana." *New York Times* 1 Dec. 2011: n. pag. *Times Topics*. Web. 5 Dec. 2011. <http://topics.nytimes.com/top/reference/timestopics/subjects/m/marijuana/index.html?scp=2&sq=Proposition%2019&st=cse>.

Nicholson, Stephen P. "The Political Environment and Ballot Proposition Awareness." *American Journal of Political Science* 47.3 (2003): 403-410. *JSTOR*. Web. 5 Dec. 2011.

Prior, Markus. "News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout." *Midwest Political Science Association* 49.3 (2005): 577-592. *JSTOR*. Web. 5 Dec. 2011. <http://www.jstor.org/stable/3647733 .>.

Rosenberg, Scott. *Say Everything: How Blogging Began, What It's Becoming, and Why It Matters*. New York: Crown Publishers, 2009. Print.

Data Sources

http://field.com/fieldpoll/propositions.html

http://www.google.com/blogsearch

http://www.nytimes.com

http://www.latimes.com

http://www.usatoday.com

http://www.sfgate.com