Identifying likely voters in pre-election surveys

Scott Keeter *Senior Survey Advisor, Pew Research Center*
Ruth Igielnik *Research Associate, Pew Research Center*

American Association for Public Opinion Research
Webinar
March 23, 2016

Learning objectives

- Understand **why likely voter models** are necessary in election polling
- List a variety of **sampling methods**, including **voter file options** and why they can be useful for election research
- Identify **typical questions included** in likely voter models
- Understand the **types of likely voter models** typically used, including probabilistic and deterministic (cut-off) methods

The problem, simply stated: Not everyone votes

Democrats are usually less likely than Republicans to vote, but often equally likely to *say* they will

The challenge

"Likely voter models are asked to produce a model of a population that does not yet exist at the time the poll is conducted, the future electorate."

Different elections, different considerations

- Different elections feature different electorates:
  - Presidential-year vs. off-year
  - general vs. primary
- We want a method that works in high-turnout and in low-turnout elections
- Important to note if a particular election:
  - Has lots of early or absentee voting
  - Is drawing more or less interest than usual
- Wild cards:
  - mobilization, demobilization, new barriers to voting

Case study

Two waves of the American Trends Panel conducted by Pew Research Center in a midterm election
Pre-Election: Sep. 9 – Oct. 3, 2014
Post-Election: Nov. 17 – Dec. 15, 2014
N = 2,424 U.S. adults who responded to both waves of the panel, were registered to vote, and were matched to a commercial voter file

House vote choice among registered, verified voters in 2014 election

	Republican	Democrat	Other	Don't know/ refused	NET
Pre-election vote choice among…					
…all registered voters	38	42	6	14	D +4
…verified 2014 voters	44	41	4	10	R+3

March 23, 2016 www.pewresearch.org 7

Why election polls sometimes go wrong

- **Biased samples** that include an incorrect proportion of each candidate's supporters
- **Change in voter preferences** between the time of the poll and the election
- **Incorrect forecasts** about who will vote (why we need likely voter models)

March 23, 2016 www.pewresearch.org 8

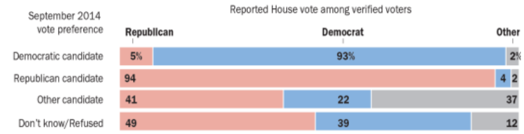
Unbiased sample

	Republican	Democrat	Other	Don't know/ refused	NET
Pre-election vote choice among…					
…all registered voters	38	42	6	14	D +4
…verified 2014 voters	44	41	4	10	R+3
Post-election vote choice among verified voters	51	45	4	-	R +6
2014 election results	51	46	3	-	R +5

March 23, 2016 www.pewresearch.org 9

There was very little switching, but late deciders broke for Republicans

Change in vote choice among panelists, pre- to post-election



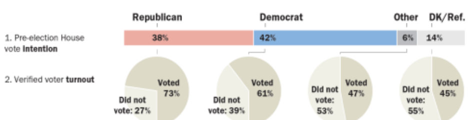
Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file.

PEW RESEARCH CENTER

Republicans turned out to vote at a higher rate than Democrats

Vote intention and turnout

Republican candidates benefited from higher turnout by their supporters. Among pre-election respondents who supported Republican candidates, 73% turned out to vote; among those who supported Democratic candidates, 61% turned out to vote.



Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter file.

PEW RESEARCH CENTER

Three big choices:

What kind of *sample*?
What *measures* to include?
How to *model* the data?

Sampling

Types of samples

- General public samples
  - RDD samples -- cellphones and landlines
  - Online samples – probability and nonprobability
- Registration based samples (RBS)
  - Commonly known as “the voter file”
  - List of all registered voters from every state; often stitched together by vendors
  - Used by many campaign pollsters
  - Contain information about past voting history for each individual, along with modeled partisanship and political engagement (also may include demographic and lifestyle information)

Some voter file vendors

Democratic
Catalist
TargetSmart

Republican
TargetPoint
DataTrust
i360

Non-partisan
Labels and Lists (L2)
Aristotle
NationBuilder
Individual states

Limitations of voter files

- Many are missed by voter files:
  - People **not listed in voter files**
  - At least 11% of the adult citizenry is not listed
    - Disproportionately affects blacks, Hispanics and the highly mobile
  - Little digital fingerprint
  - New voters and **new registrants**
- Even among people who are listed in the voter file, information may not be up to date or accurate
  - **In our case study, <u>16% say they voted in 2014 but have no record of voting</u>** (on voter file)
  - Some of this is **over-reporting**, but its also plausible that the **absence of a record is an error**
- Differs by state

1. Jackman, Simon and Bradley Spahn. 2015b. "Unlisted in America." Unpublished paper. Accessed Dec. 22, 2015, at https://www.dropbox.com/s/bv5z1tyv9q422aw/Jackman%20Spahn%20-%20Unlisted%20in%20America.pdf?dl=0

Measures for identifying likely voters

Perry scale method

- Developed in the 1950s and 1960s by Paul Perry of Gallup
- Derives a likely voter index from questions aimed at measuring:
  - Political engagement
  - Experience with voting
  - Past vote
  - And intention to vote
- Cutoff at expected turnout
- Commonly-used method

Example of questions in Perry scale method

How much thought have you given to the coming November election?
***Quite a lot, some**, only a little, none*

Have you ever voted in your precinct or election district?
***Yes**, no*

Would you say you follow what's going on in government and public affairs…
***Most of the time, some of the time**, only now and then, hardly at all?*

How often would you say you vote?
***Always, nearly always**, part of the time, seldom*

How likely are you to vote in the general election this November?
***Definitely will vote, probably will vote**, probably will not vote, definitely will not vote*

In the 2012 presidential election between Barack Obama and Mitt Romney, did things come up that kept you from voting, or did you happen to vote?
***Yes, voted**; no*

Please rate your chance of voting in November on a scale of 10 to 1.
*0-8, **9, 10***

*the categories that give a respondent a point in the Perry scale, discussed in the following section, are in bold

Measures of vote intention are important but insufficient

Share of registered voters		% who are verified 2014 voters	Margin
	Likelihood of voting		
70	**Definitely will vote**	77	D +2
20	**Probably will vote**	36	D +18
7	Probably will not vote	13	D +24
2	Definitely will not vote	15	D +64
100			
	Chance of voting		
75	**9-10 (higher likelihood)**	75	D +2
11	7-8	34	D +13
5	5-6	27	D +39
4	3-4	16	D +10
4	1-2 (lower likelihood)	8	D +12
100			

The most engaged voters lean Republican

Share of registered voters		% who are verified 2014 voters	Margin
	Thought about election		
29	**Quite a lot**	82	R +17
40	**Some**	62	D +12
21	Only a little	55	D +10
10	None	29	D +40
100			
	Follow government and public affairs		
40	**Most of the time**	78	R +14
37	**Some of the time**	62	D +15
16	Only now and then	43	D +29
8	Hardly at all	30	D +9
100			

Self-reports of past voting predict future voting, but are not perfect

Share of registered voters		% who are verified 2014 voters	Margin
	2012 vote (self-reported)		
87	**Voted**	70	D +4
12	Did not vote	17	D +8
1	Too young to vote	27	D +55
100			
	How often vote		
45	**Always**	82	R +2
33	**Nearly always**	59	D +8
13	Part of the time	34	D +9
9	Seldom	17	D +37
100			
	Ever voted in precinct or election district		
83	**Yes**	70	D +2
17	No	26	D +21

Combining the measures to form the Perry scale

Score on scale	Share of all verified voters	% who are verified voters in each group
7	63%	83%
6	15	63
5	10	59
4	4	34
3	4	41
2	2	23
1	1	13
0	1	11
	100	

22% of verified voters scored between 0 and 5

83% of those who scored a 7 actually voted

Source: 2014 American Trends Panel September and November waves. Based on registered voters who participated in both waves and were matched to a national voter.

Approaches to identifying likely voters

Deterministic	
WHAT: • **Ask of all** adults/registered voters • Identify likely voters with a **series of questions used in a scale.** • Only include **most likely to vote** (top of scale) according to expected turnout.	WHO: • Gallup • Pew Research Center ABC/Washington Post • LA Times • Quinnipiac Poll • Time • IPSOS (also uses weighting)
Probabilistic	
Model as you sample	

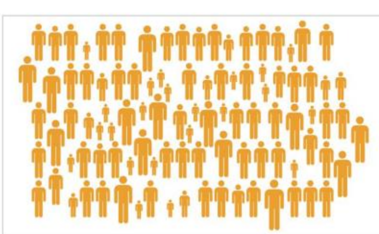
Approaches to identifying likely voters

Deterministic methods

Approaches to identifying likely voters

Deterministic	
Probabilistic	
WHAT: • Same questions, **used as a weight.** • Include everyone, **weighted to their probability of voting.**	WHO: • CBS/New York Times • Marist
Model as you sample	

Approaches to identifying likely voters

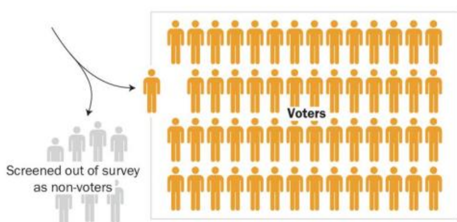


Probabilistic methods

Approaches to identifying likely voters

Deterministic	
Probabilistic	
Model as you sample	
WHAT: • Usually rely on a combination of **past vote from voter file** and **self-reported vote** on a sliding scale • **Sample using past vote** • Construct sample to **reflect probability of voting**. • Goal to increase efficiency	WHO: Primarily partisan pollsters including: • Anzalone Liszt Grove Research • Democracy Corps • the Field Poll

Approaches to identifying likely voters

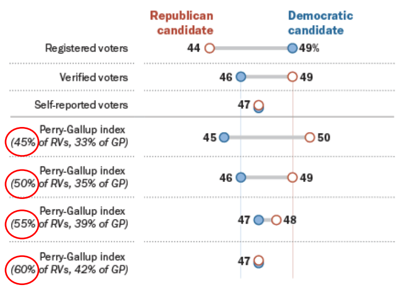


Model as you sample

Deterministic (cut-off) methods

- One example: the Perry scale
- Categorize each survey respondent as a likely voter or nonvoter
- Uses a threshold or “cutoff” that matches the predicted rate of voter turnout in the election
- The turnout forecast matters

Turnout forecast matters



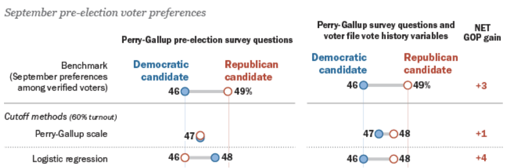
Probabilistic or modeling approaches

- Same set of questions as Perry scale
- Calculates a probability of voting for each survey respondent
- Variety of methods including logistic regression, random forest
- Can include records of past turnout from the voter file
- Models could be applied to future elections to test viability

Voter file measures of past voting behavior predict future voting

Share of registered voters		% who are verified 2014 voters	Margin
	Verified past vote		
78	Voted in 2012	75	D +2
22	No record	21	D +18
100			
55	Voted in 2010	84	R +6
45	No record	36	D +19
100			

Use past voting history (if you have it) to strengthen estimates of the likelihood of voting



March 23, 2016 www.pewresearch.org 34

Machine learning approaches

March 23, 2016 www.pewresearch.org 35

Machine learning models

- Random Forest
  - Able to sort through large numbers of variables to find patterns
  - Typical decision tree analysis identifies various ways of splitting a dataset into separate paths or branches, based on options for each variable
  - Random forest uses large number of decision trees to split the data into similar groups
  - Probabilistic model – produces a predicted probability for each respondent

March 23, 2016 www.pewresearch.org 36

Optimizing your models with limited space

•Machine learning models using combinations of questions show promise if you have limited space

	Republican	Democrat	Margin
Overall machine learning model (random forest)	48	46	R +2
Ground truth	49	46	R +3
Registered voters in survey	38	42	D +4
Plan to vote in the general election	46	48	D +2
Voted in 2012	45	49	D +4
Plan to vote, voted in 2012	46	48	D +2
Plan to vote, voted in 2012, voted in your precinct before	47	48	D +2
Plan to vote, voted in 2012, how often do you vote in elections	47	48	D +2
Plan to vote, voted in 2012, thought about the election	47	47	EVEN
Plan to vote, voted in 2012, follow government and politics	47	47	EVEN

Applying models to future elections

- Tricks to apply models to future election data
  - Use regression coefficients to create a predictive equation – apply to new dataset
  - Using statistical software makes it simpler!
    - STATA: Run regression on old dataset, use "predict" function right away with new dataset
    - SPSS: Specify in regression parameter under SELECT which dataset you're predicting onto
    - R: Use "predict" function and specify "new data" parameter

Take-home considerations

Can't take people at their word about voting
- Republicans more likely than Democrats to vote
- Over-reporting of intention, but under-reporting too

Use multiple questions in multiple dimensions
- Intention, past behavior, engagement
- The specific questions may not matter

Various models (deterministic vs. probabilistic) can work
- Cutoff methods adjust to different election circumstances
- Probabilistic methods use all available data
- But harder to calibrate and implement

Voter files can be very useful
- Valuable in low-turnout elections
- Useful for targeting
- But issues of coverage remain

Additional resources

Ansolabehere, Stephen, and Eitan Hersh. 2012. "Validation: What Big Data Reveal About Survey Misreporting and the Real Electorate." *Political Analysis* 20:437-59.

Bernstein, Robert, Anita Chadha, and Robert Montjoy. 2001. "Overreporting Voting: Why it Happens and Why it Matters. *Public Opinion Quarterly* 65:22-44.

Berent, M. K., Krosnick, J. A., & Lupia, A. (2011). The quality of government records and "over-estimation" of registration and turnout in surveys: Lessons from the 2008 ANES Panel Study's registration and turnout validation exercises (Working Paper No. nes012554). Ann Arbor, MI: American National Election Studies. Retrieved from http://www.electionstudies.org/resources/papers/nes012554.pdf

Burden, Barry. 1997. "Deterministic and Probabilistic Voting Models." *American Journal of Political Science* 41:1150-69.

Colasanto, Diane and Jay A. Mattlin. 1987. "Evaluation of Gallup's Methodology for Predicting Likelihood of Voting." Paper prepared for presentation at the 1987 Joint Statistical meetings, San Francisco, Aug. 17, 1987.

Crespi, Irving. 1988. *Pre-election Polling: Sources of Accuracy and Error*. New York: Russell Sage Foundation.

Dimock, Michael, Scott Keeter, Mark Schulman and Carolyn Miller. 2001. "Screening for Likely Voters in Pre-election Surveys." Presented at the Annual Meeting of the American Association for Public Opinion Research, Montreal, May 17-20.

Additional resources

Freedman, Paul and Ken Goldstein. 1996. "Building a Probable Electorate from Pre-election Polls: A Two-Stage Approach." *Public Opinion Quarterly* 60:574-87.

Jackman, Simon and Bradley Spahn. 2015b. "Unlisted in America." Unpublished paper. Accessed Dec. 22, 2015, at https://www.dropbox.com/s/bv5z1tyv9q422aw/Jackman%20Spahn%20-%20Unlisted%20in%20America.pdf?dl=0

Perry, Paul. 1960. "Election Survey Procedures of the Gallup Poll." *Public Opinion Quarterly* 24:531-42.

Perry, Paul. 1979. "Certain Problems in Election Survey Methodology." *Public Opinion Quarterly* 43:312-25.

Pew Research Center. 2016. "Can Likely Voter Models Be Improved? Evidence from the 2014 U.S. House Elections." Available at http://www.pewresearch.org/2016/01/07/can-likely-voter-models-be-improved/

Rogers, Todd and Masahiko Aida. 2013. "Vote Self-Prediction Hardly Predicts Who Will Vote, and Is (Misleadingly) Unbiased." *American Politics Research,* 503-528.

Silver, Brian D., Barbara A. Anderson, and Paul R. Abramson. 1986. "Who Overreports Voting?" *American Political Science Review* 80:613-24.

Zukin, Cliff. 2004. "Sources of Variation in Published Election Polling: A Primer." American Association for Public Opinion Research. Available at http://aapor.org/pdfs/varsource.pdf

Questions?

Scott Keeter:
skeeter@PewResearch.org

Ruth Igielnik:
rigielnik@PewResearch.org