Polling

Gov 1347: Election Analytics

Kiara Hernandez

September 22, 2022

Harvard University

But first, a review of 3 crucial concepts

Multivariate (multiple) regression

Univariate regression ($\underbrace{Y}_{PV} = \alpha + \beta \underbrace{X_1}_{GDP}$) $\rightsquigarrow$

multivariate regression ($\underbrace{Y}_{PV} = \alpha + \beta_1 \underbrace{X_1}_{GDP} + \beta_2 \underbrace{X_2}_{approval}$)

- assumption: linear additivity between IVs ($\beta_1 X_1 + \beta_2 X_2$)
- estimation: same OLS procedure (minimizing the sum of squared errors)
- better in-sample fit metric: $R^2_{adjusted} = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$ for p IVs
- adding new IVs will change parameter estimates
    - Intuitively, β_1 is the `unique` effect of X_1
    - only two rare cases where $\hat{\beta} = \hat{\beta}_1$: 1. $\beta_2 = 0$

      2. $r_{X_1,X_2} = 0$
    - signs and magnitudes of correction $\hat{\beta}_2$ and r_{X_1,X_2} $\rightsquigarrow$ signs and magnitudes of correction $\hat{\beta}_1 - \hat{\beta}$

Interactions of IVs

Sociopolitical outcomes (DVs) are often predicted by **interactions** of IVs:

- intersectionality: (`race x gender`)
  - black women face more discrimination than black men and white women do combined
  - the effect of X_1 (race) can vary depending on the value of X_2 (gender)
- extension 2:
  - (`incumbent party x economic performance`)
  - (`year x economic performance`)
- extension 3: (`party x unemployment`)
  - in state *s*, incumbent party benefits if Dem, hurt if Rep (why?)

Overfitting

> *A common criticism of fundamentals models is that they are extremely easy to **over-fit**—the statistical term for deriving equations that provide a close match to historical data, but break down when used to predict the future. To avoid this risk, we borrow two techniques from the world of machine learning, with appropriately inscrutable names: elastic-net regularisation*** *and **leave-one-out cross-validation**.*

(Morris 2020a)

Explicit tension between **in-sample fit** (“close match to historical data”) and **out-of-sample performance**:

- Cross-validation
- Elastic-net regularisation: Parsimony of a model** reduces out-of-sample error
- Bottom-line: an $R^2 > 0.9$ might actually be bad for prediction!

Today's agenda

Can we predict election outcomes using polling data?

1. **A brief overview of polls and pollsters**
   - How do polls work?
   - Pros and cons
2. **Quantitatively describing the polls:**
   - How do polls fluctuate across state, time, and year?
   - How early can the polls predict election results?
3. **Improve our 2020 forecast using polling data**
   - How to resolve with fundamentals model(s) from last week?

Polls and Pollsters

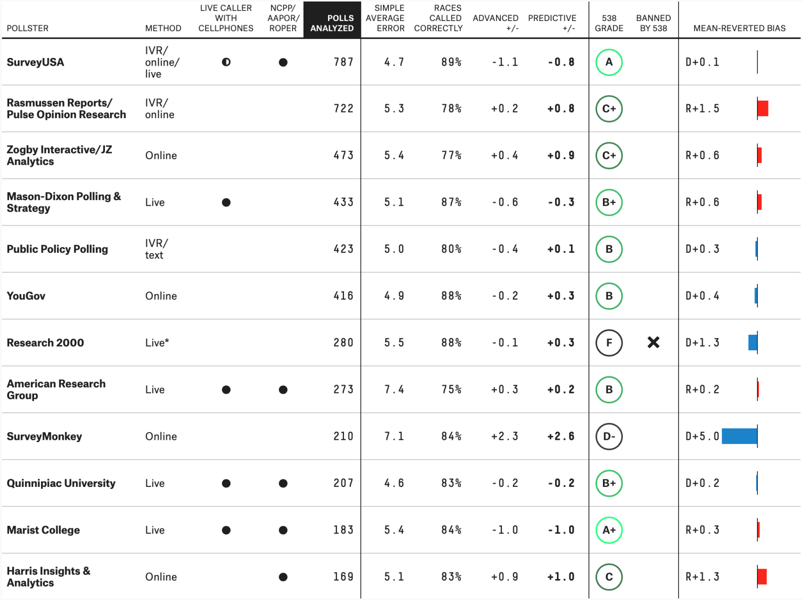
Pollsters

What do they do?

Organizations that conduct public opinion research by:

1. Designing a questionnaire
   - vote choice
   - generic ballot
   - presidential approval
2. Contacting a sample (often opaque)
   - phone vs. internet
   - random vs. non-random
3. Ask repeatedly over time
   - panel (rare)
   - cross-section
4. Weight responses to “look like population’ ’ (often opaque)
   - choice of variables
   - choice of models
5. Interpret and report

Some pollsters you might know



POLLSTER	METHOD	LIVE CALLER WITH CELLPHONES	NCPP/ AAPOR/ ROPER	POLLS ANALYZED	SIMPLE AVERAGE ERROR	RACES CALLED CORRECTLY	ADVANCED +/-	PREDICTIVE +/-	538 GRADE	BANNED BY 538	MEAN-REVERTED BIAS
SurveyUSA	IVR/ online/ live	◐	●	787	4.7	89%	-1.1	**-0.8**	A		D+0.1
Rasmussen Reports/ Pulse Opinion Research	IVR/ online			722	5.3	78%	+0.2	**+0.8**	C+		R+1.5
Zogby Interactive/JZ Analytics	Online			473	5.4	77%	+0.4	**+0.9**	C+		R+0.6
Mason-Dixon Polling & Strategy	Live	●		433	5.1	87%	-0.6	**-0.3**	B+		R+0.6
Public Policy Polling	IVR/ text			423	5.0	80%	-0.4	**+0.1**	B		D+0.3
YouGov	Online			416	4.9	88%	-0.2	**+0.3**	B		D+0.4
Research 2000	Live*			280	5.5	88%	-0.1	**+0.3**	F	✗	D+1.3
American Research Group	Live	●	●	273	7.4	75%	+0.3	**+0.2**	B		R+0.2
SurveyMonkey	Online			210	7.1	84%	+2.3	**+2.6**	D-		D+5.0
Quinnipiac University	Live	●	●	207	4.6	83%	-0.2	**-0.2**	B+		D+0.2
Marist College	Live	●	●	183	5.4	84%	-1.0	**-1.0**	A+		R+0.3
Harris Insights & Analytics	Online		●	169	5.1	83%	+0.9	**+1.0**	C		R+1.3

Live (live telephone interviews including cell calls), Live (not including cell calls), Online (internet, text, etc.), IVR (interactive voice response)*

Why shouldn't we trust the polls?

- Non-response
  - Ex: In 2016, "less-educated' ' whites systematically opted out of polls
  - Ex: Convention bounce, enthusiasm gap
- Respondent dishonesty
  - 23% of the Harvard-hosted CES (CCES) takers lied about voting!
- Respondent confusion
  - "Do you [support President Bush's / favor or oppose] decision sending additional troops to Iraq?"
- Reponses not weighted "correctly' '
  - NYT: In 2016, four professional pollsters independently weighted a Florida poll. Only one predicted Trump would win (+1%).
- Polls misinterpreted by pundits, commentators, voters ("madness of crowds' ')

Why should we trust the polls?

- Wisdom of crowds (Galton 1907)
- Wisdom of *aggregators* of crowds
  - Ex: FiveThirtyEight, Real Clear Politics, The Economist
- Can adjust averages across time and across polls (Silver 2016) # update this?
- Interpretation of vote-choice polls is straight-forward
- Predicts past elections (how well, though?)

Describing the relationship between polls and election outcomes

How do polls fluctuate across time?

Let’s visualize poll averages by party from the most recent midterm election:

```
library(tidyverse)
poll_df <- read_csv("polls_df.csv")
```

```
poll_df %>%
```

```
  filter(year == 2018) %>%
```

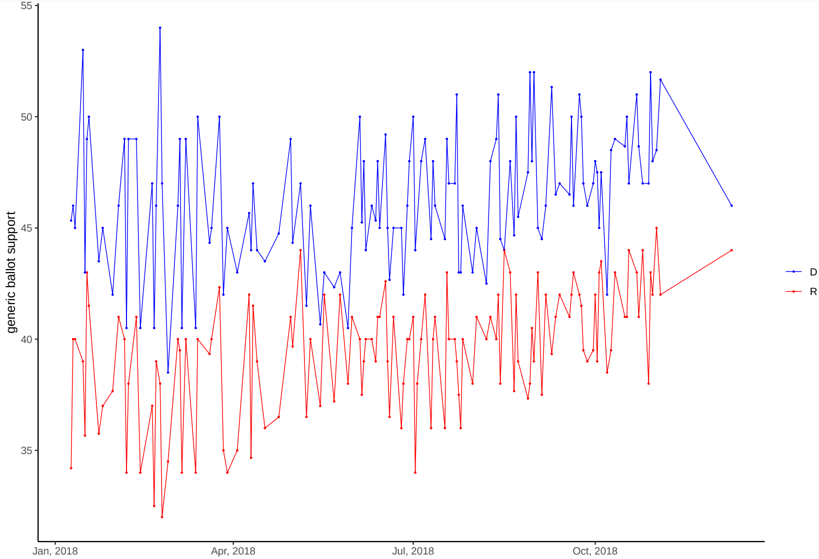
```
  ggplot(aes(x = poll_date, y = support, color = party)) +
```

```
    geom_point(size = 1) +
    geom_line() +
```

```
    scale_x_date(date_labels = "%b, %Y") +
    scale_color_manual(values = c("blue","red"), name = "") +
    ylab("polling approval average on date") + xlab("") +
```

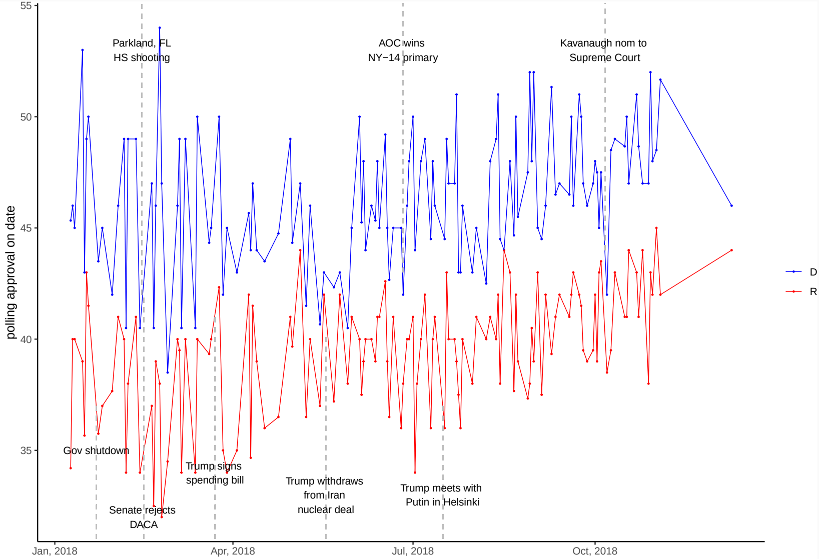
How do polls fluctuate across time?

Poll averages in 2016:



How do polls fluctuate across time?

What events were "game-changers" in 2018?



How do polls fluctuate across time?

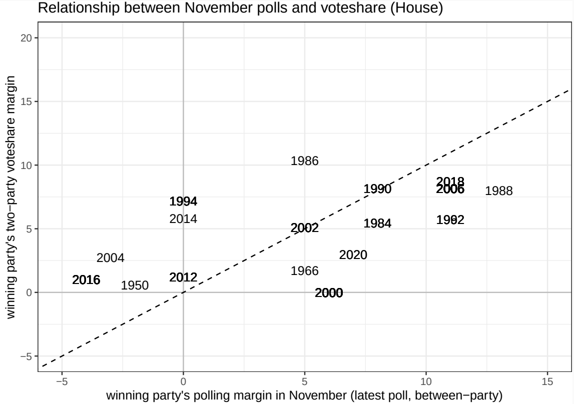
- Convention and debate bumps.
- Polls fluctuate after some “game-changers”.
- Polls fluctuate (or don’t) *despite* “game-changers”.
- **Momentum** is a phrase used $\geq$ 60 times a day by media outlet in election season, but...
  - Denter and Sisak, *Journal of Public Economics* (2015): in close races, poll equilibrium often shifts after a bump to one candidate
  - FiveThirtyEight (2010): weak evidence of positive serial correlation in polls (some evidence of negative serial correlation!)

Does the November poll predict the election?

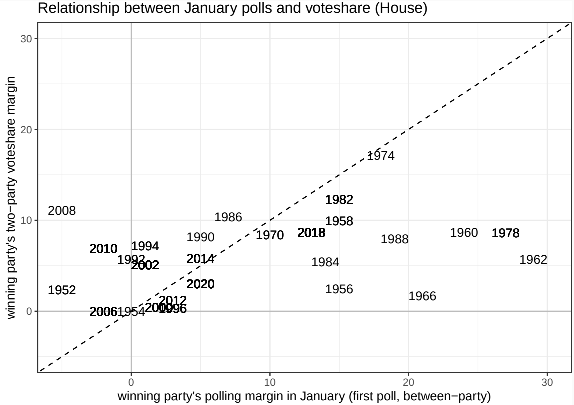
Correlation between November poll margin and two-party PV margin is:

0.3700875

Does the November poll predict the election?



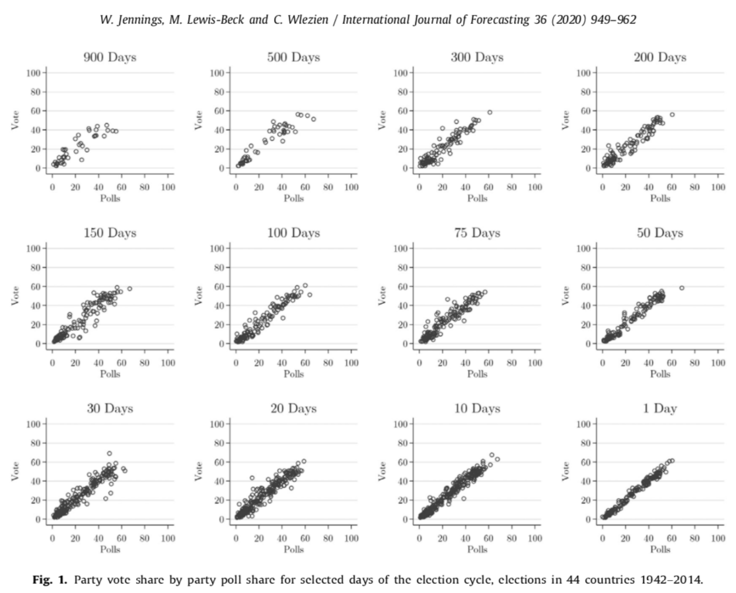
Does the January poll predict the election?



Correlation between January poll margin and two-party PV margin is:

0.6555578

Do polls get more predictive over time?



W. Jennings, M. Lewis-Beck and C. Wlezien / International Journal of Forecasting 36 (2020) 949–962

Fig. 1. Party vote share by party poll share for selected days of the election cycle, elections in 44 countries 1942–2014.

Predicting 2020 using polls

Predicting 2020 using polls

(Interactive Session in `R Studio`)

Predicting 2020 using polls

What we've learned:

- One-sided model of incumbent $\rightsquigarrow$ two-sided model of incumbent and challenger
    - using `pv2p` usually means $\hat{Y}_{inc} + \hat{Y}_{chl} > 100\%$
    - can be fit in a single `lm` using interactions
- **Classification accuracy:** $\frac{\text{number of correct predictions}}{\text{number of predictions}}$
- Multiple regression is useful, but not only way to combine IVs
- **Weighted ensembles** $\rightsquigarrow$ flexible to combine separate models, e.g.:
    - weight equally (Graefe 2020)
    - weight on days til election (pollsters vs. Gelman & King (1993))
    - weight on R^2 (Silver 2022)
    - weight on cross-validation error (this is called "Super Learning")
    - weight on human priors
- The most cutting edge methods, to date, combine fundamentals and polls using probabilistic (Bayesian) models, e.g. Lauderdale, Linzer (2015) $\rightsquigarrow$ stay tuned!

Update your forecasting model

Moving forward, you will put out a forecast every week that builds on previous weeks. This week, add in polling data in whatever way you see fit (could be guided by the extensions). How does your model change? Is it a better model (as determined by in-sample and out-of-sample tests)? Worse? Next week, you can adjust your model again depending on what you find.

Blog Extensions

1. **What Do Forecasters Do?** Based on what you've learned about fundamentals and poll-based forecasts, (1) briefly summarize Silver (2022) and Morris (2020a): (https://projects.economist.com/us-2020-forecast/house) [1] and (2) compare and contrast their approaches. In your opinion, which of the two is the better approach?
2. **Incorporating Pollster Quality:** Consider 2018/2020 pollster ratings (on `GitHub`) from FiveThirtyEight. (1) How much variation is there in pollster quality? (2) Using tools and knowledge you've gained so far, build a model (possibly an ensemble) using individual polls from 2018 (`538_generic_2018.csv`) and 2022 (`538_generic_poll_2022.csv`). How does your model compare to the models this week in lab?
3. **Incorporating district-level polls** How do district-level polls differ from national level polls? Using careful model evaluation techniques and considering possible choices of weighted ensembles, build a predictive model for 2022 using current cycle district-level polls. Remember that you can choose two-party voteshare or seatshare as your outcome variable. How you build the model is up to you. You can combine district and national-level polls. You can use only current cycle district-level polls as your predictor. You can combine historical national-level polls from past cycles with current cycle district-level polls. You can weight district and national-level differently.

[1] Don't worry about the technical details (regularization/shrinkage/Bayesian modeling)