The Air War

Gov 1347: Election Analytics

Soubhik Barari, Sun Young Park

October 6, 2020

Harvard University

Today's agenda

- **Introduction to probabilistic models for election forecasting**
    - what problem it solves
    - brief intro to binomial logistic regression
    - simulating a distribution of election outcomes in Pennsylvania 2020
- **Advertising data mini-hackathon (20 minutes)**
- **Simulating advertising effects in 2020**
    - using existing models of ad effects
- **Suggested blog extensions**
    - build your own model of ad effects
    - build your own probabilistic models
    - what's the deal with social media?

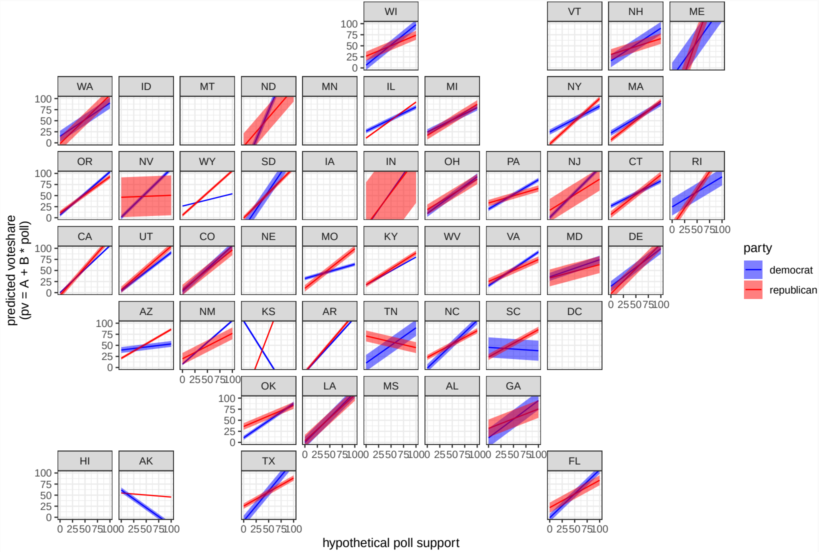
Introduction to probabilistic models

One major problem with linear regression

When we fit a linear regression model $Y = \alpha + \beta X$, there are no restrictions on Y. What's wrong with that?

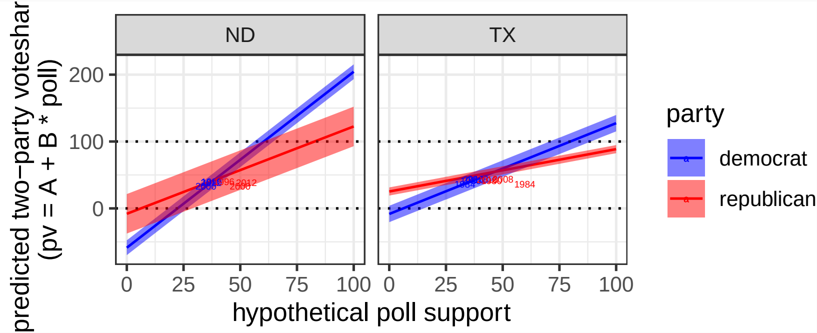
- $\rightsquigarrow$ With one of our fundamentals models, Y (Trump 2020 PV) had a prediction interval lower bound < 0 (**out of support**).
- **This often occurs when we are extrapolating but also when there is sparse data (e.g. when we fit a linear regression model on state-level polls).**

Poll-only state-level linear regression predictions



Q: What's wrong with this map?

Poll-only state-level linear regression predictions



Solution: probabilistic models

Linear regression: outcome can be any value in a continuous range (-∞, +∞)

$$\%DemPV_{state} = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k \quad \text{or}$$

and modeled as

$$DemPV_{state} = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k,$$

but the true outcome is bounded to (0, 100) or (0, VEP_{state}) . . .

Binomial logistic regression: election outcome for Democrats is finite draw of voters from the voter-eligible population (VEP_{state}) turning out to vote Democrat (a binomial process) modeled as

$$Pr(\underbrace{\text{Vote for Dem}_{state,i}}_{voter\ \ i\ \ in\ \ state}) = f(\alpha + \beta_1 x_1 + \ldots + \beta_k x_k)$$

$$= \frac{exp(\alpha + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + exp(\alpha + \beta_1 x_1 + \ldots + \beta_k x_k)} \text{ (for i = 1, ..., } VEP_{state}\text{)}$$

where link function f (inverse logistic function) bounds (-∞, +∞) to (0, 100)

Example of a probabilistic model: binomial logistic regression

Supposing we have `x` (a single IV), `y` (a DV) as a % which is computed from `y = draws/popl`:

	Linear regression	**Binomial logistic regression (binomial logit)**
link function	$f(\alpha + \beta x) = \alpha + \beta x$	$f(\alpha + \beta x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$
link function name	identity	inverse logistic function
link function output	predicted outcome	predicted probability of one draw
`R` code	`lm(y∼x)`	`glm(cbind(draws, popl-draws)∼x, family=binomial)`
fitting intuition	"do OLS to find coefficients that minimize $\sum (y - \hat{y})^2$"	"find coefficients where fitted draw probabilities $f(\hat{\alpha} + \hat{\beta} x)$ best predict observed `draws` for all x"
prediction intuition	"plug in x_{new} and get (i) predicted outcome $\hat{y}_{new} = \hat{\alpha} + \hat{\beta} x_{new}$ and (ii) prediction interval $\hat{y}_{new} \pm 1.96 \times \text{se}(\hat{y}_{new})$"	"plug in x_{new} and get (i) predicted probability of one draw, $f(\hat{\alpha} + \hat{\beta} x_{new})$; also plug in `popl` to get (ii) predicted expected number of draws, $\widehat{\texttt{draws}}$, and (iii) predicted distribution of draws from repeated binomial process simulations"

Poll-only state-level binomial logit

State-level dataset merged with VEP data:

```
vep_df <- read_csv("vep_1980-2016.csv")
poll_pvstate_vep_df <- pvstate_df %>%
  mutate(D_pv = D/total) %>%
  inner_join(pollstate_df %>% filter(weeks_left == 5)) %>%
  left_join(vep_df)

ND_D <- poll_pvstate_vep_df %>%
  filter(state=="North Dakota", party=="democrat")
```

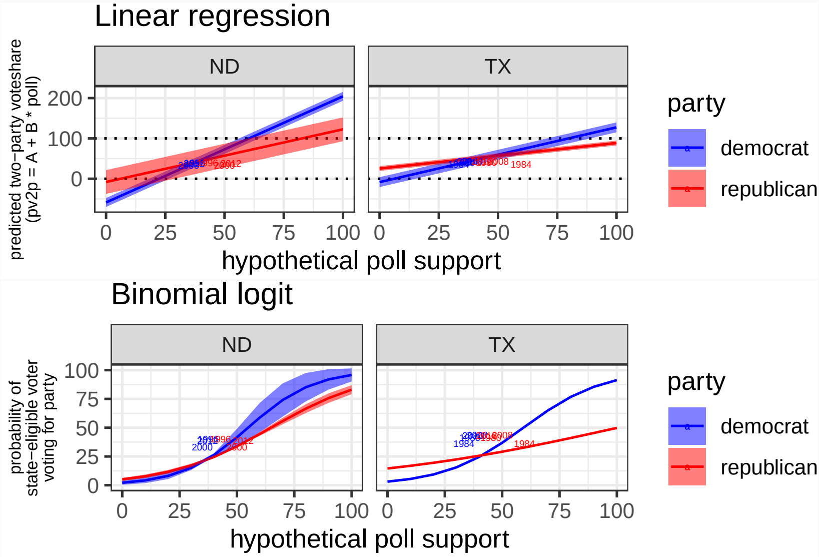
Linear regression model for North Dakota race (Dem side):

```
ND_D_lm <- lm(D_pv ~ avg_poll, ND_D)
```

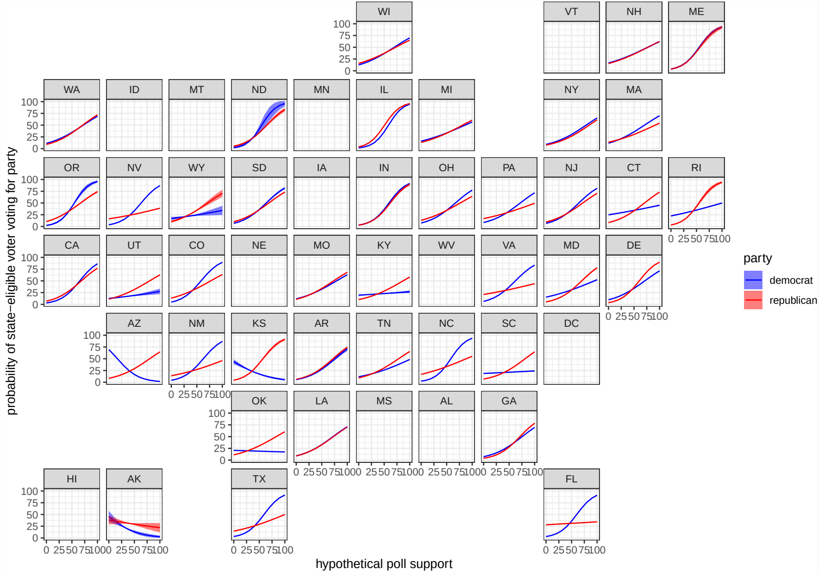
Binomial logit model for North Dakota race (Dem side):

```
ND_D_glm <- glm(cbind(D, VEP-D) ~ avg_poll, ND_D,
                family = binomial)
```

Comparison of poll-only state-level predictions



Poll-only state-level binomial logit predictions (all states)



Simulating a distribution of election results

Instead of (i) a probability for a single D voter or (ii) single expected number of D voters from `popl`, $\widehat{\texttt{draws}}$, we can predict a (iii) distribution of draws from binomial process on that `popl`.

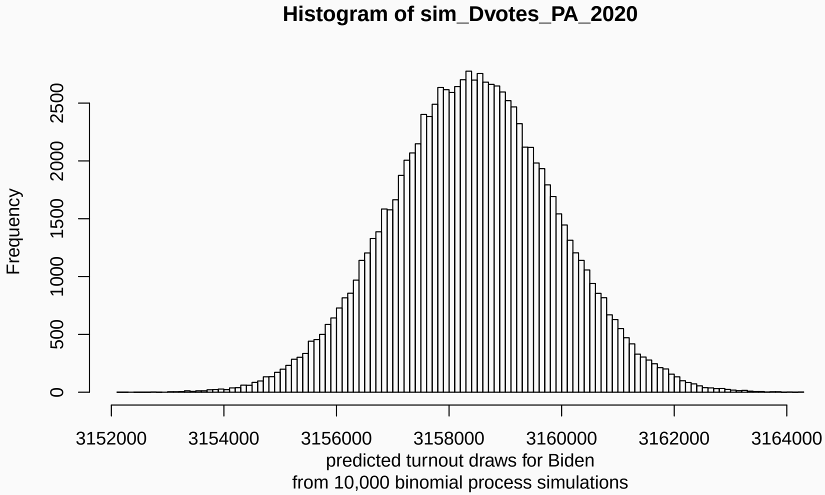
```r
## Get relevant data
VEP_PA_2020 <- as.integer(vep_df$VEP[vep_df$state == "Pennsylvania" & vep_df$year == 2016])

PA_R <- poll_pvstate_vep_df %>% filter(state=="Pennsylvania", party=="republican")
PA_D <- poll_pvstate_vep_df %>% filter(state=="Pennsylvania", party=="democrat")
```

```r
## Fit D and R models
PA_R_glm <- glm(cbind(R, VEP-R) ~ avg_poll, PA_R, family = binomial)
PA_D_glm <- glm(cbind(D, VEP-D) ~ avg_poll, PA_D, family = binomial)
```

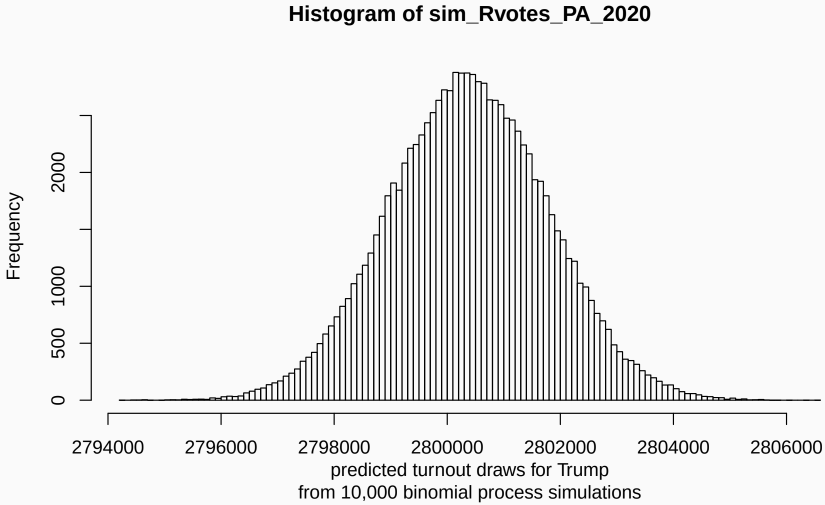
```r
## Get predicted draw probabilities for D and R
prob_Rvote_PA_2020 <- predict(PA_R_glm, newdata = data.frame(avg_poll=44.5), type="response")[[1]]
prob_Dvote_PA_2020 <- predict(PA_D_glm, newdata = data.frame(avg_poll=50), type="response")[[1]]
```

```r
## Get predicted distribution of draws from the population
sim_Rvotes_PA_2020 <- rbinom(n = 100000, size = VEP_PA_2020, prob = prob_Rvote_PA_2020)
sim_Dvotes_PA_2020 <- rbinom(n = 100000, size = VEP_PA_2020, prob = prob_Dvote_PA_2020)
```

Simulating a distribution of election results: Biden PA PV

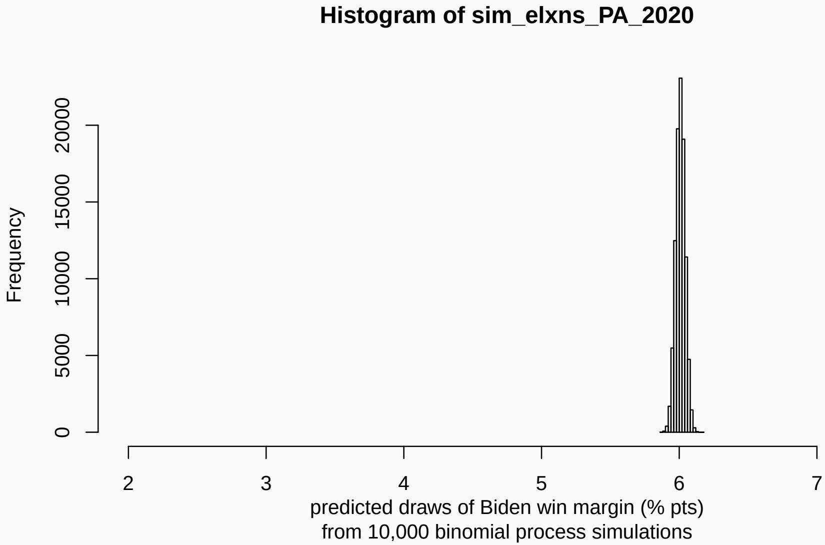


Simulating a distribution of election results: Trump PA PV



Simulating a distribution of election results: Biden win margin

```
sim_elxns_PA_2020 <- ((sim_Dvotes_PA_2020-sim_Rvotes_PA_2020)/(sim_Dvotes_PA_2020+sim_Rvotes_PA_2020))*100
```



Q: Does this seem plausible? How could we improve this?

Summary of probabilistic models

- Explicitly capture a random or probabilistic process of the world
  - ex: some draw of voters from VEP turning out

- Models like binomial logit (**generalized linear models**) use a link function to bound the outcome to a probability value
  - link functions like the inverse logistic function allow us to **non-linearly** predict DV from IVs (solving another problem of linear regression)

- Workflow: estimate the parameters of a probabilistic model ⇝ obtain distributions from repeated simulations of probabilistic process
  - ex: in binomial logit, we repeatedly draw voters from a binomial process based on predicted probability of one voter turning out Dem
  - $\sim$ how The Economist simulates elections

- Diagnostics: can still use out-of-sample evaluation tools; see glossary and `had.co.nz/notes/modelling/logistic-regression.html` for other diagnostics.

Advertising data mini-hackathon

Presidential ad campaigns from 2000-2012

`ad_creative_2000-2012.csv`: data related to the design and content ("creative") of candidate/party-run ads.

`creative`	`party`	`ad_issue`	`cycle`	`ad_purpose`	`ad_tone`
ad name	D/R	WMP-coded issue	election year	WMP-coded purpose	WMP-coded tone

`ad_campaigns_2000-2012.csv`: day-by-date ad spendings and airings in each state by candidates/parties/PACs in each campaign.

...	`sponsor`	`state`	`creative`	`n_markets`	`n_stations`	`total_cost`	`after_primary`
	PAC/candidate sponsor		ad name	# ads aired across DMAs	# ads aired across stations	total ad spend ($)	date after party primary?

`ads_2020.csv`: partial and temporally aggregate state-level data around total spendings and airings by Biden/Trump in 2020.

`state`	`period_startdate`	`period_enddate`	`biden_airings`	`trump_airings`	`total_airings`	`total_cost`
			all D ads aired	all R ads aired	all D/R ads aired	total cost (D and R)

Advertising data mini-hackathon (20 minutes)

It is the year 2024 and you and a few of your Gov 1347 friends have landed a lucrative data science internship for the DNC. In 20 minutes, you all have a meeting with the chief campaign strategist of [your favorite soon-to-be presidential candidate] who wants answers to the following questions:

1. What are some trends in the tone and purpose of Democrat ads for each party?
2. What are some trends in the issues mentioned in ads by each party?
3. How much money do winning challenger candidates spend on advertising and when during the campaign do they spend most of it?
4. Obama '08 sure was an amazing run ... how much $ did his team spend on battleground states and which did they win?

Read and merge together `ad_creative_2000-2012.csv` with `ad_campaigns_2000-2012.csv` and answer these 4 questions with your team. If you're running short on time, guesstimate the remaining answers.

Simulating advertising effects in 2020

A hypothetical ad war in PA

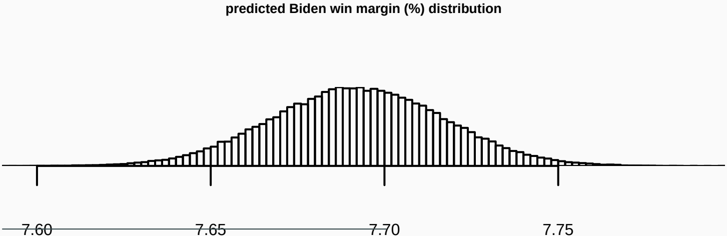
From readings: 1 **gross point rating (GRP)**[1] ad buy in PA is $\approx$ \$300.

Let's take the models in our readings at face value:

- $1,000$ GRPs buys a 5($\pm$1.5 s.e.) pts shift in vote (Gerber et al.)
- $1,000$ GRPs buys a 7.5($\pm$2.5 s.e.) pts shift in vote (Huber et al.)

Let's take the current FiveThirtyEight aggregate PA polls as voter turnout probabilities in Binomial process (49% Biden, 42% Trump):

```
sim_Dvotes_PA_2020 <- rbinom(n = 100000, size = VEP_PA_2020, prob = 0.49)
sim_Rvotes_PA_2020 <- rbinom(n = 100000, size = VEP_PA_2020, prob = 0.42)
sim_elxns_PA_2020 <- (sim_Dvotes_PA_2020-sim_Rvotes_PA_2020)/(sim_Dvotes_PA_2020+sim_Rvotes_PA_2020)*100
```



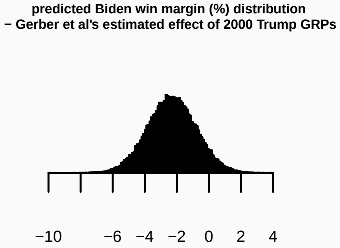
---

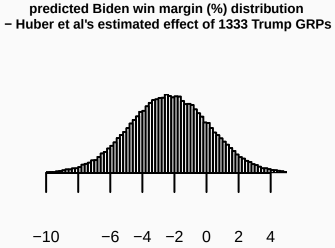
[1] GRP = % of reachable audience × number of airings

A hypothetical ad war in PA

For $\approx 2\%$ win margin, Trump would need to gain $\approx 10\%$ or 2000/1333 more GRPs ($\approx \$0.4$-$0.6$ million) than Biden a week before election:

```
sim_elxns_PA_2020_shift.a <- sim_elxns_PA_2020 - rnorm(100000, 10, 1.5)
sim_elxns_PA_2020_shift.b <- sim_elxns_PA_2020 - rnorm(100000, 10, 2.5)
```

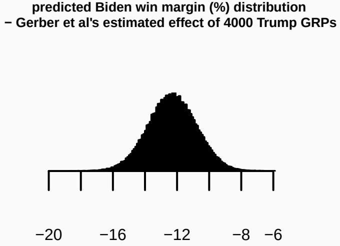

predicted Biden win margin (%) distribution − Gerber et al's estimated effect of 2000 Trump GRPs

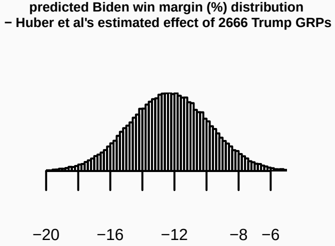

predicted Biden win margin (%) distribution − Huber et al's estimated effect of 1333 Trump GRPs

A hypothetical ad war in PA

For ≈ 12% win margin, Trump would need to gain ≈ 20% or 4000/2666 more GRPs (≈ \$0.8-1.2 million) than Biden a week before election:

```r
sim_elxns_PA_2020_shift.a <- sim_elxns_PA_2020 - rnorm(100000, 20, 1.5)
sim_elxns_PA_2020_shift.b <- sim_elxns_PA_2020 - rnorm(100000, 20, 2.5)
```

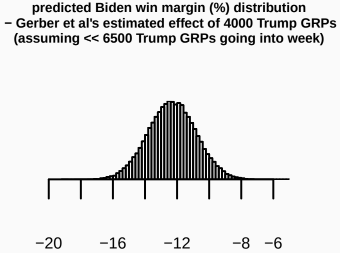


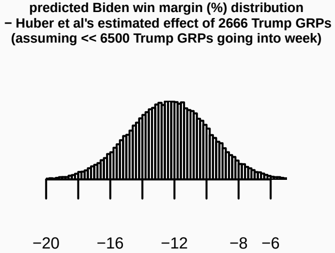


A hypothetical ad war in PA: dimishing returns of GRPs

For ≈ 12% win margin, Trump would need to gain ≈ 20% or 4000/2666 more GRPs (≈ $0.8-1.2 million) than Biden a week before election but only if Trump had bought << 6500 GRPs to begin with according to Huber et al.:

```r
sim_elxns_PA_2020_shift.a <- sim_elxns_PA_2020 - rnorm(100000, 20, 1.5)
sim_elxns_PA_2020_shift.b <- sim_elxns_PA_2020 - rnorm(100000, 20, 2.5)
```


predicted Biden win margin (%) distribution
− Gerber et al's estimated effect of 4000 Trump GRPs
(assuming << 6500 Trump GRPs going into week)

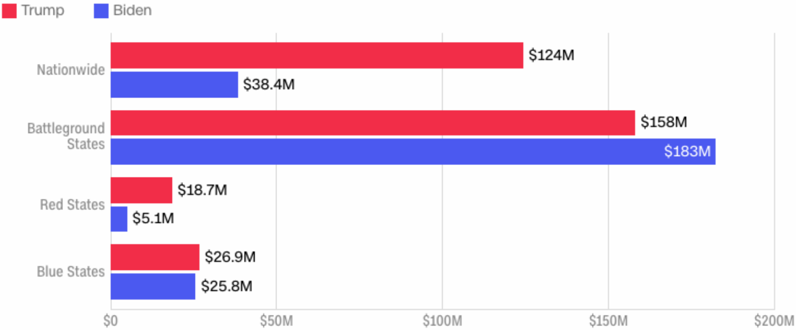

predicted Biden win margin (%) distribution
− Huber et al's estimated effect of 2666 Trump GRPs
(assuming << 6500 Trump GRPs going into week)

Q: Is this plausible?

The 2020 ad campaigns so far

Biden has so far outspent Trump in battleground states

With the help of outside groups, Biden caught up to Trump's one-year head start and has outspent Trump in the 10 battleground states as of Aug. 31. However, the president still has the advantage not just in red states, but a slight spending lead in blue states as well.



Note: Figures cover all digital, TV and radio ad spending both by the candidates' campaigns and by affiliated PACs and super PACs in the general election cycle, which for President Donald Trump spans the period from Jan. 1, 2019 through Aug. 31, 2020. Biden's spending reflects Feb. 11, 2020 through Aug. 31, 2020.

Source: Kantar Media/CMAG
Graphic: Christopher Hickey, CNN

Q: How would you incorporate ads into your predictive model?

Blog Extensions (optional)

Using Partial 2020 Ads Data. Using the data from this section (and incorporating useful data from previous weeks) fit a model and predict Biden and Trump's 2020 voteshare in each state given partial data on ad spending up to Sep 27 (given in `ads_2020.csv`)[2]. How does your prediction of PA 2020 compare to our simulation of effects from Huber et al. (2007) and Gerber et al. (2011)? What are the limitations of your model?

Probabilistic Simulation of State-Level Races. Extend the binomial regression-based simulation we did of the Pennsylvania 2020 race to all 2020 races based on the most recent poll numbers for Biden and Trump. Make a `geofacet` map of the distribution of your predictions. Do they make sense? Speculate as to why or why not.

Social Media. How much do campaigns spend on social media ads? Does social media influence election outcomes? How will it influence 2020?

. . . and lots of rich descriptive opportunities for this week!

---

[2] Hint: you will either have to choose your model based on what data is available or use informed guesses about each campaign's spending in the relevant time period.

Famous Presidential Political Ads

- Daisy - LBJ, 1964
- Morning in America - Reagan, 1984
- Willie Horton - H.W. Bush, 1988
- Smoking Man - Herman Cain, 2011
- Luis Bracamontes - Donald Trump, 2018