The Ground Game

Gov 1347: Election Analytics

Kiara Hernandez

October 13, 2022

Harvard University

Today's agenda

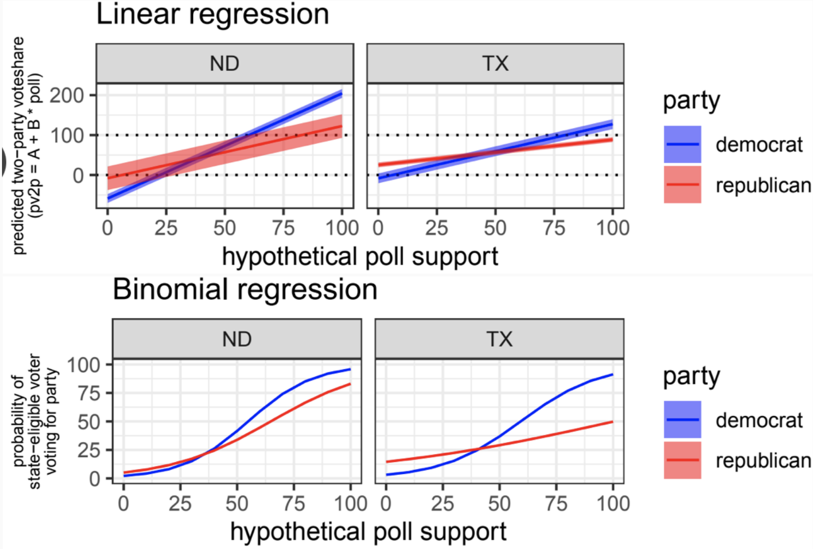
- **Introduction to probabilistic models for election forecasting**
    - what problem it solves
    - brief intro to binomial logistic regression
    - simulating a distribution of election outcomes in Pennsylvania District 01 for 2022
- **Prediction: why use turnout? Thinking through campaigning - persuasion vs. mobilization**

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?

- $\rightsquigarrow$ It is possible to have a prediction interval lower bound < 0 (**out of support**).
- **This can occur when we are extrapolating but also when there is sparse data**
  (e.g. when we fit a linear regression model on district-level polls).

Poll-only district-level linear regression vs. binomial logistic predictions



Q: What's wrong with these plots?

Solution: probabilistic models

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

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

and modeled as

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

but the true outcome is bounded to (0, 100) or (0, $CVAP_{district}$) . . .

Binomial logistic regression: election outcome for Democrats is finite draw of voters from the citizen voting age population ($CVAP_{district}$) turning out to vote Democrat (a binomial process) modeled as

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

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

where link function f (inverse logistic function) bounds (-∞, +∞) to (0, 1)

Example of a probabilistic model: binomial logistic regression

Supposing we have `x` (a single IV), `y` (a DV) as Dem's popular vote share (%):

	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, cvap-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`"
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 `CVAP` to get (ii) predicted expected number of draws, $\widehat{\texttt{draws}}$, $\rightsquigarrow \frac{\widehat{\texttt{draws}}}{\texttt{CVAP}}$ and (iii) predicted distribution of draws from repeated binomial process simulations"

Simulating a distribution of election results: Extension #2

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

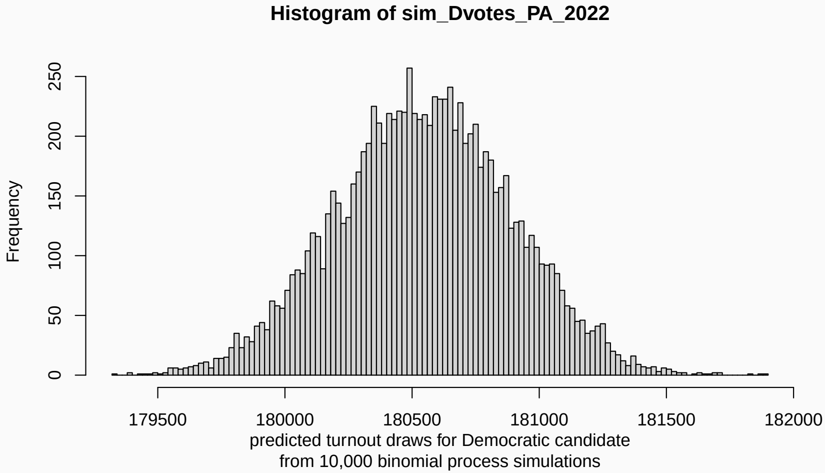
```
##
## 2018 2020
##  760  371
```

```
## Fit D and R models
PA_R_glm <- glm(cbind(RepVotes, cvap-RepVotes) ~ REP, PA01,
                family = binomial)
PA_D_glm <- glm(cbind(DemVotes, cvap-DemVotes) ~ DEM, PA01,
                family = binomial)
```

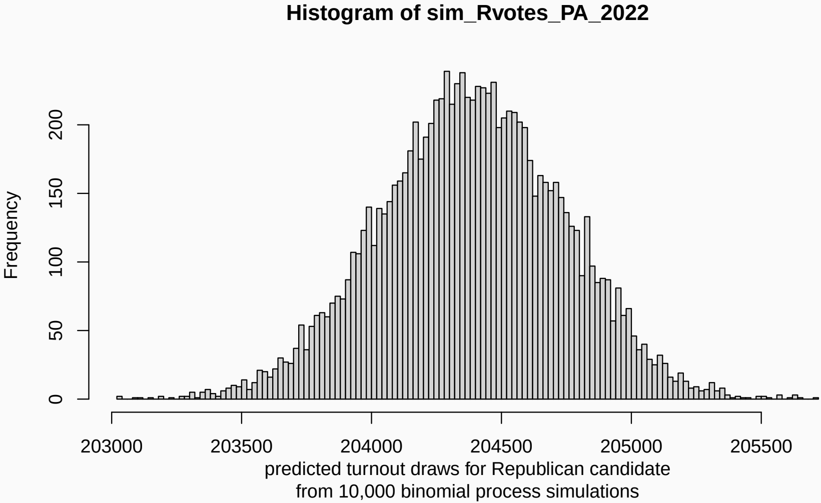
```
## Get predicted draw probabilities for D and R
prob_Rvote_PA_2022 <- predict(PA_R_glm, newdata =
                                data.frame(REP=44.5),
                              type="response")[[1]]

prob_Dvote_PA_2022 <- predict(PA_D_glm, newdata =
                                data.frame(DEM=50),
                              type="response")[[1]]
```

```
## Get predicted distribution of draws from the population
sim_Rvotes_PA_2022 <- rbinom(n = 10000, size = CVAP_PA_2022, prob = prob_Rvote_PA_2022)
sim_Dvotes_PA_2022 <- rbinom(n = 10000, size = CVAP_PA_2022, prob = prob_Dvote_PA_2022)
```

Simulating a distribution of election results: Democrat PA01 PV

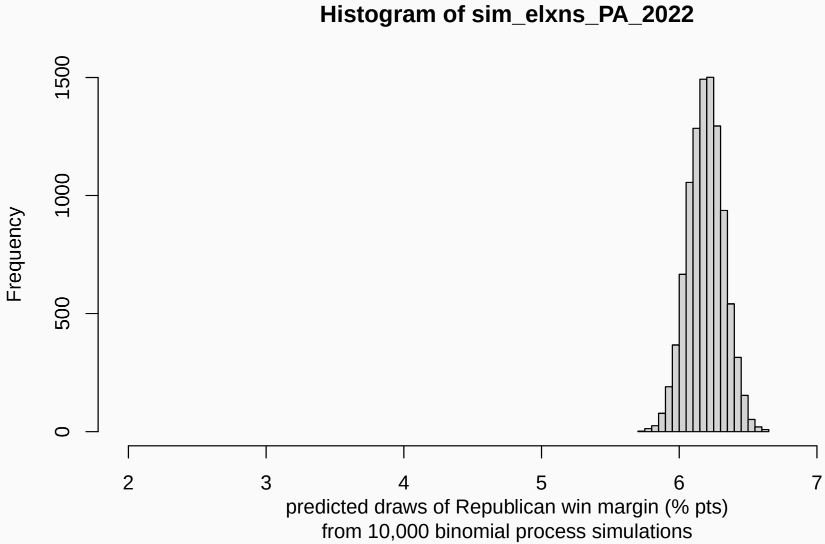


Simulating a distribution of election results: Republican PA01 PV



Simulating a distribution of election results: Republican win margin in PA01

```
sim_elxns_PA_2022 <- ((sim_Rvotes_PA_2022-sim_Dvotes_PA_2022)/(sim_Dvotes_PA_2022+sim_Rvotes_PA_2022))*100
```



Reading and interpreting GLMs

```
# linear regression for PA01
PA01_lm <- lm(DemVotesMajorPercent ~ DEM, data = PA01)
summary(PA01_lm)
```

```
##
## Call:
## lm(formula = DemVotesMajorPercent ~ DEM, data = PA01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6532 -2.0630 -0.2078  2.4366  3.4172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.4897     7.4261   7.876 1.04e-06 ***
## DEM          -0.2650     0.1562  -1.697     0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.58 on 15 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.1051
## F-statistic: 2.879 on 1 and 15 DF,  p-value: 0.1104
```

Reading and interpreting GLMs

```r
# binomial logit
PA01_glm <- glm(cbind(DemVotes, cvap-DemVotes) ~
                  DEM, PA01, family = binomial)
summary(PA01_glm)
```

```
##
## Call:
## glm(formula = cbind(DemVotes, cvap - DemVotes) ~ DEM, family = binomial,
##     data = PA01)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -56.357  -40.450    2.937   33.752   59.935
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.3114193  0.0083238 -157.55   <2e-16 ***
## DEM          0.0125754  0.0001748   71.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 32280  on 16  degrees of freedom
## Residual deviance: 27116  on 15  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 27350
##
## Number of Fisher Scoring iterations: 3
```

Reading and interpreting GLM predictions

```r
# predict lm
prob_PA01_lm <- predict(PA01_lm, newdata =
                              data.frame(DEM=46))
prob_PA01_lm
```

```
##        1
## 46.29815
```

Reading and interpreting GLM predictions

```
# predict glm
prob_PA01_glm <- predict(PA01_glm, newdata =
                           data.frame(DEM=46), type="response")[
prob_PA01_glm
```

```
## [1] 0.3245478
```

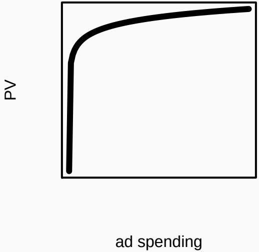
Summary of probabilistic models

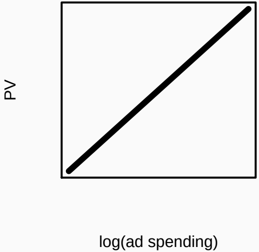
- Explicitly capture a random or probabilistic process of the world
    - ex: some draw of voters from CVAP turning out
- Models like binomial logit (**generalized linear models**) use a link function to bound the outcome to a probability value
    - link functions like the inverse logistic function allow us to **non-linearly** predict DV from IVs (solving another problem of linear regression)
- Workflow: estimate the parameters of a probabilistic model $\rightsquigarrow$ 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 `had.co.nz/notes/modelling/logistic-regression.html` for other diagnostics.

Blog tip 2

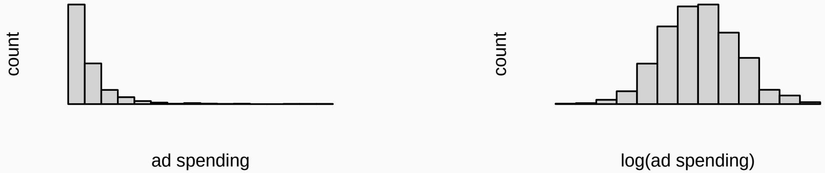
Recall our conversation with Prof. Vavreck → **Should `log()` a "skewed" variable like ad spend. Why?**

1. modeling: diminishing returns of $1 ⇝ log-transformation linearizes the relationship.





2. description: when most ad spends small, few ad spends huge ⇝ log-transformation makes it easier to see/count these outliers.



Blog tip 3

District s has too little poll data to fit a model. What do I do?

- no model: predict PV as literal 2022 poll value, but report the out-of-sample error of raw polls
- non-poll model: use other data (e.g. local economy) for district, but report the out-of-sample error of model
- pooled model: rather than running district by district regression, use district-level poll model that's fit *across* all districts (more on this today!)
- no polls whatsoever: use previous election results, generic ballot polls, polls from districts with similar characteristics, as we've done in past weeks

Turnout

Our weeks on campaigns are trying to understand how voters respond to ads and on-the-ground campaigning efforts. From our readings, we know that campaigns try to do two things:
* (1) mobilize - turning people out to vote * (2) persuade - convincing people out to vote for a particular candidate/party

First, let's look at turnout at the district-level to identify any interesting patterns across time and geography.
Then, let's turn back to last week's ad data to see whether there is a relationship between turnout and ad spends.

How do we calculate turnout?

*$turnout_{district_i} = \frac{totalvotes}{CVAP}$ where *totalvotes* is the number of two-party votes cast in a given district in a given year and CVAP is the citizen voting-age population in a given district in a given year. Note the differences between CVAP, voting-age population (VAP), voting eligible population (VEP)

- CVAP = total population that is age 18+ and a citizen
- VAP = total population that is age 18+
- VEP = all U.S. citizens age 18+, who are not excluded from voter eligibility due to criminal status (felony convictions, incarceration, or parole)

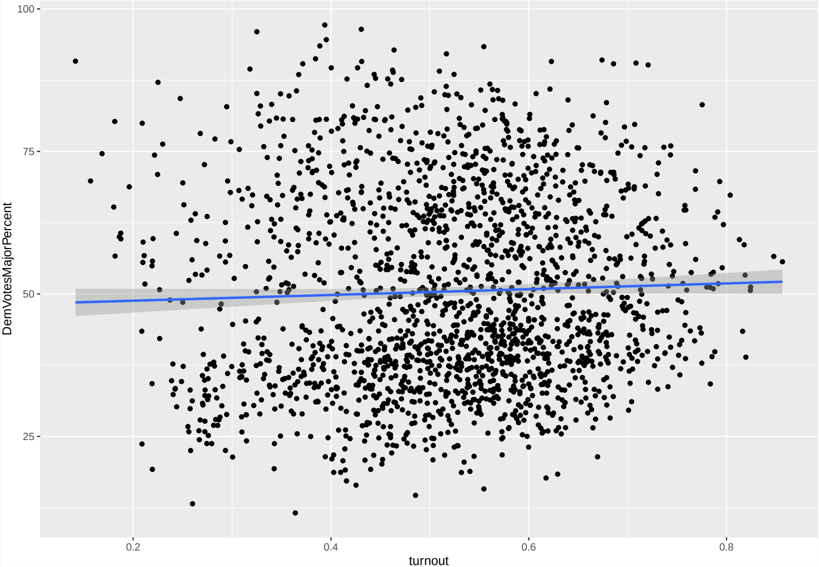
*The data we're working with comes from the American Community Survey's 5-Year Estimates. The 5-year estimates are "period" estimates that represent data collected over a period of time. The primary advantage of using multiyear estimates is the increased statistical reliability of the data for less populated areas and small population subgroups.

*This means that when a CVAP estimate exists for 2018, we can use that estimate for all years between 2012-2018.

Turnout and Democratic voteshare

```r
# visualize
ggplot(dist_pv_cvap_closed, aes(x = turnout, y = DemVotesMajorPe
  geom_point() +
  stat_smooth(method = "lm")
```

Turnout and Democratic voteshare



What other relationships related to turnout would you be interested in exploring?

What other relationships related to turnout would you be interested in exploring?

- Aggregate changes in turnout over time
- District-level changes in turnout over time
- ______?
- Spend some time talking in small groups and (if we have time) beginning to explore what you come up with.

Blog Extensions

Turnout model: Incorporate turnout, incumbency and expert predictions into your district-level two-party voteshare predictions.

Close Elections: (i) Do expert predictions predict turnout? Fit a model and discuss your results, thinking specifically about whether your results provide evidence for the effect of ground campaigns on turnout.
(ii) Do ad spends predict turnout? Merge last week's WMP data with this week's data on turnout at the district level. What can we infer, if anything, about the relationship between the "air war" and voter persuasion/mobilization?

Probabilistic Simulation of District-Level Races. Update your working forecasting model from one that is based on linear regression to one that is modeled as a GLM.

Extend the binomial regression-based simulation we did of the Pennsylvania 2022 race to all 2022 races based on the most recent poll numbers for the Democratic and Republican candidate (in districts for which district-level polling data is available). Make a `geofacet` map of the distribution of your predictions. Do they make sense? Speculate as to why or why not.