Survey Research and Design

Small Group Estimation

William Marble

October 31, 2023

Public Opinion in the Nation and in the States

- Why is public opinion important?

Public Opinion in the Nation and in the States

- Why is public opinion important?
- Do politicians care about public opinion? Why?

Public Opinion in the Nation and in the States

- Why is public opinion important?
- Do politicians care about public opinion? Why?
- *Whose* opinions matter?

Public Opinion in the Nation and in the States

- Why is public opinion important?
- Do politicians care about public opinion? Why?
- *Whose* opinions matter?

Normative Reasons to Care About Small(ish) Subgroups

- No politician (other than the president) is elected by the entire country
- House members represent $\sim$ 760,000 people $\rightsquigarrow$ what kind of representation do voters in the district want?
- What do marginalized groups want from government?
- What are they getting?

Estimating Public Opinion Nationally

Recall: for estimating a mean in the population, there are diminishing returns to sample size

Estimating Public Opinion Nationally

Recall: for estimating a mean in the population, there are diminishing returns to sample size

- The standard error has $\sqrt{N}$ in the denominator ($se(\bar{y}) = \frac{\sigma_y}{\sqrt{N}}$)
- To cut standard error in half, need 4x sample size
- Justification for relatively small sample:

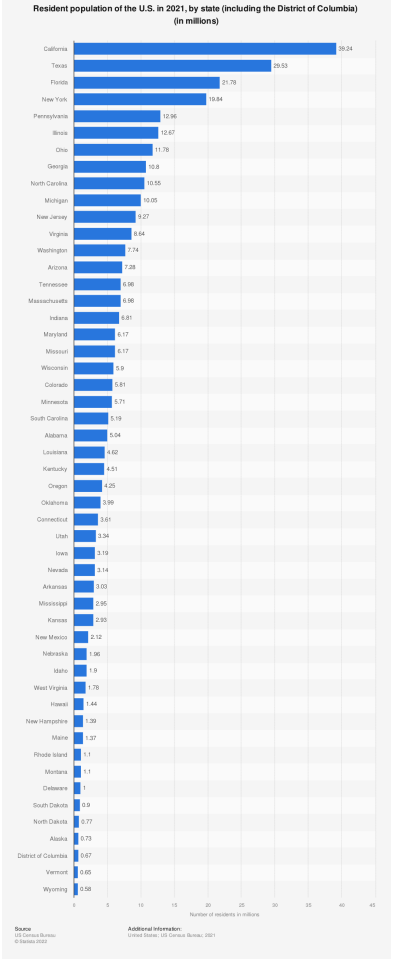
  for a binary (yes/no) question, the *maximum possible* SE under simple random sampling is $\sqrt{\frac{0.5 \times (1-0.5)}{n}}$, or 0.015 if $N = 1000$

Small Groups

Suppose we do a national survey with $N = 1,000$. How many respondents from West Virginia do we expect to get?

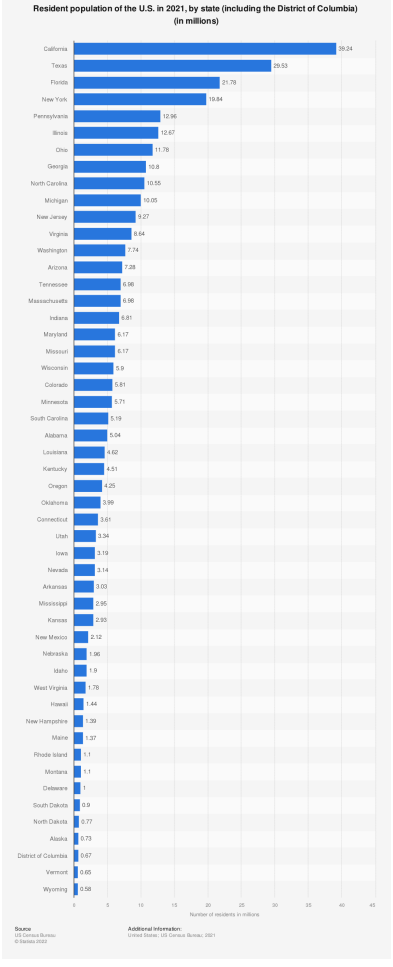
Small Groups

Suppose we do a national survey with $N = 1,000$. How many respondents from West Virginia do we expect to get?



Small Groups

Suppose we do a national survey with $N = 1,000$. How many respondents from West Virginia do we expect to get?



$\rightsquigarrow$ only 5 respondents!

Smaller Subgroups

What if we are interested in how Black women will vote in the Georgia senate race?

Smaller Subgroups

What if we are interested in how Black women will vote in the Georgia senate race?

Consider 2022 PORES poll: we started with 86,335 likely voters...

Smaller Subgroups

What if we are interested in how Black women will vote in the Georgia senate race?

Consider 2022 PORES poll: we started with 86,335 likely voters...

- ...of those, 2,790 are in GA

Smaller Subgroups

What if we are interested in how Black women will vote in the Georgia senate race?

Consider 2022 PORES poll: we started with 86,335 likely voters...

- ...of those, 2,790 are in GA
- ...of those, 1,469 are women

Smaller Subgroups

What if we are interested in how Black women will vote in the Georgia senate race?

Consider 2022 PORES poll: we started with 86,335 likely voters...

- ...of those, 2,790 are in GA
- ...of those, 1,469 are women
- ...of those, 411 are black

Smaller Subgroups

What if we are interested in how Black women will vote in the Georgia senate race?

Consider 2022 PORES poll: we started with 86,335 likely voters...

- ...of those, 2,790 are in GA
- ...of those, 1,469 are women
- ...of those, 411 are black

$N = 86k$ is a huge survey, but sample size is much smaller for important subgroups

Problems With Small Sample Sizes

- No one from a subgroup $\rightsquigarrow$ no estimate!
- Small $N \rightsquigarrow$ high levels of uncertainty
- More likely to get proportions closer to 0 or 1 than we would if we had a larger N (higher variance)
- Analogy: much more likely to get 100% heads by flipping 4 coins than by flipping 100

Problems With Small Sample Sizes

- No one from a subgroup $\rightsquigarrow$ no estimate!
- Small $N \rightsquigarrow$ high levels of uncertainty
- More likely to get proportions closer to 0 or 1 than we would if we had a larger N (higher variance)
- Analogy: much more likely to get 100% heads by flipping 4 coins than by flipping 100

Takeaway: even a reasonably sized nationally representative sample won't give us good estimates in smaller subgroups.

Potential Solutions

1. Use administrative data that covers the full population
2. Combine multiple surveys together
3. Use a statistical model to generate an estimate

Administrative Data

- If surveys aren't best suited to the task, maybe we should turn to administrative data that covers the whole population
- E.g., we could use election results, IRS data, data from business payroll records, etc.

Election Returns

Maybe we could just use election results instead of public opinion surveys?

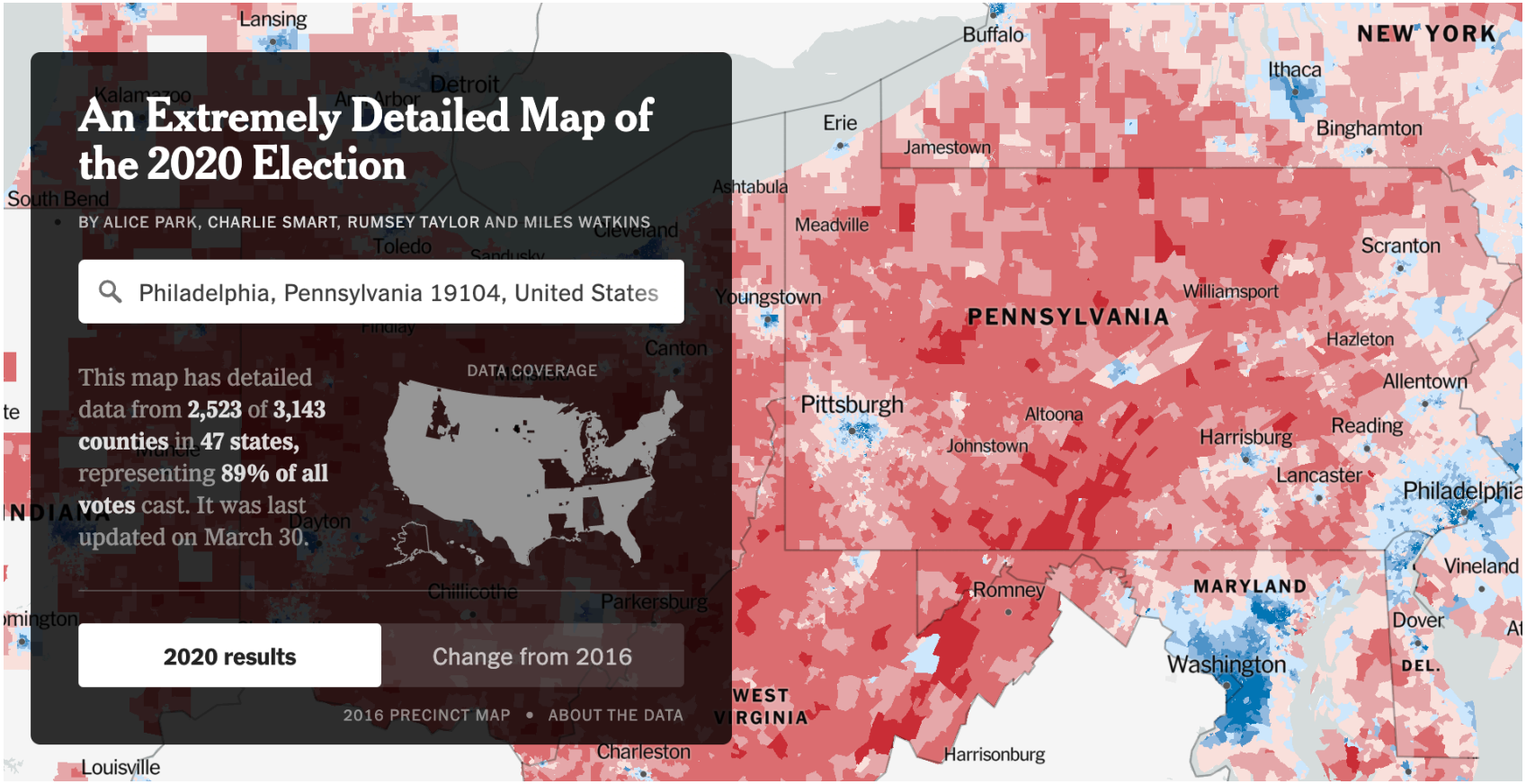
Election Returns

Maybe we could just use election results instead of public opinion surveys?

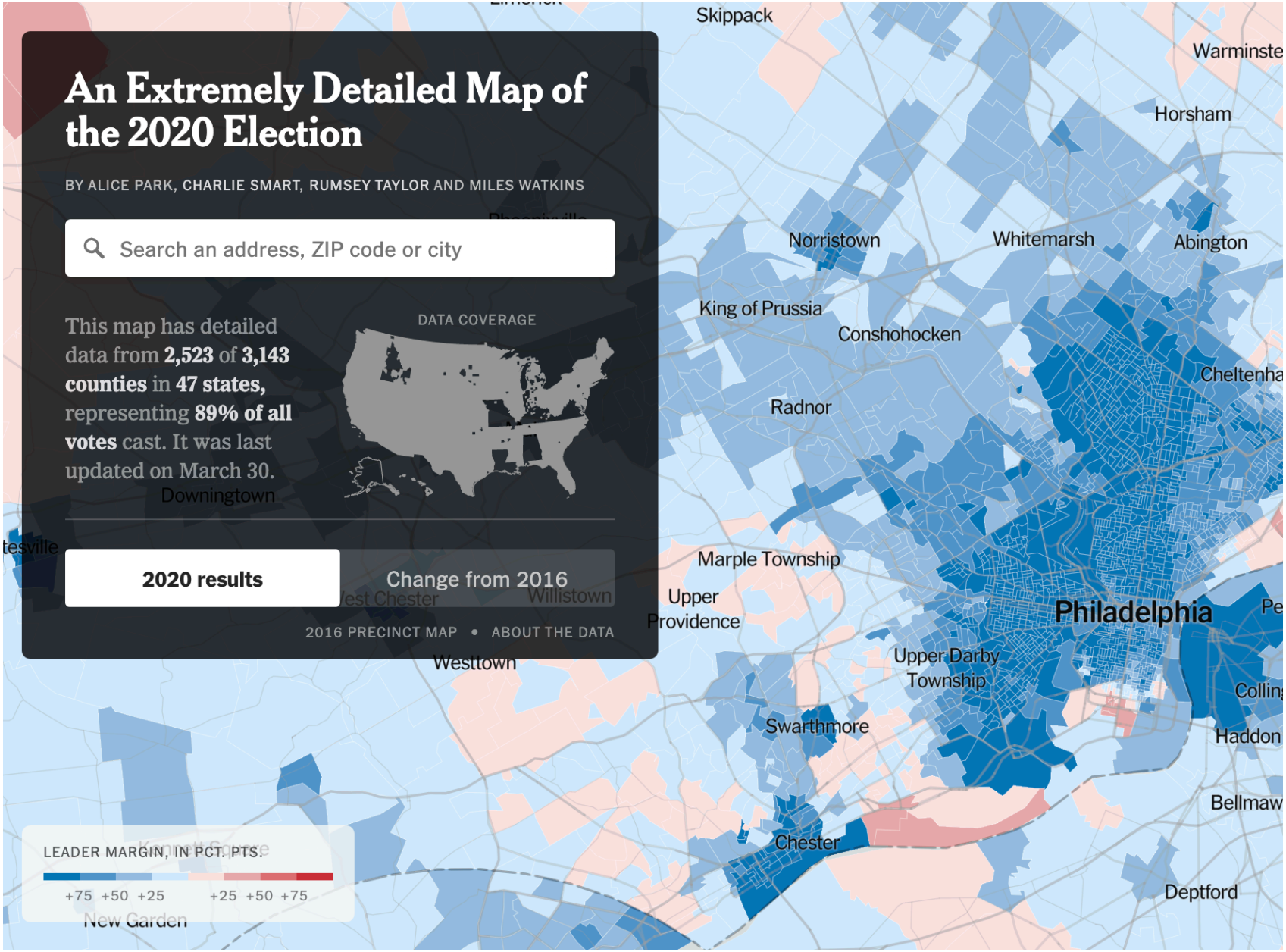
Advantages

- Ultimately, electing representatives are the main way people influence government
- There's no sampling error, nonresponse bias, or other sources of survey error
- We have fine-grained data on election results at small geographic levels

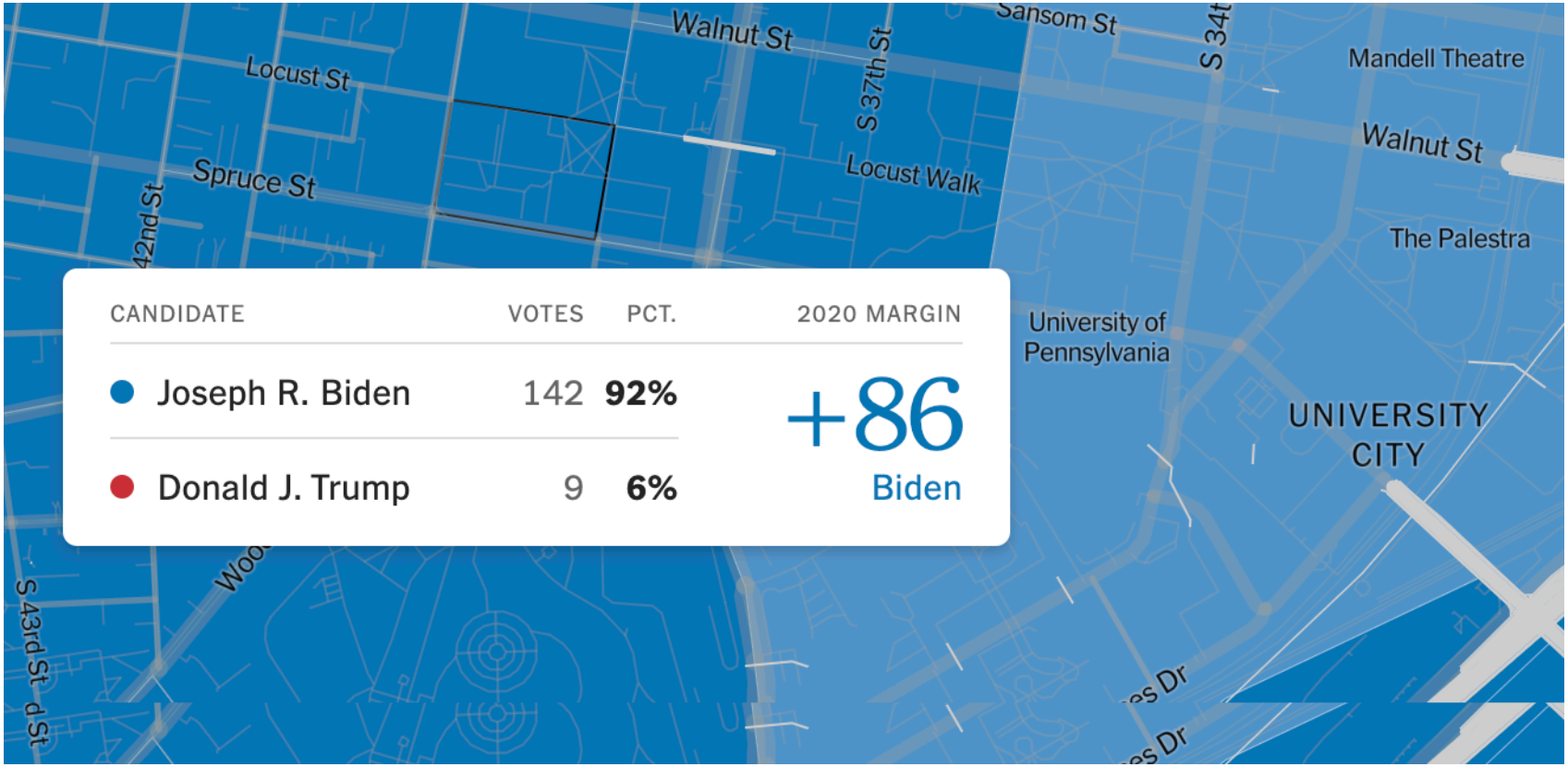
Election Returns



Election Returns



Election Returns



Limits of Election Returns

What limits do election results have for understanding public opinion?

Limits of Election Returns

What limits do election results have for understanding public opinion?

- Aggregation: can only examine groups defined by geography, when there are normative reasons to care about other groups
- Turnout: not everyone who lives in an area turns out to vote
- Issues: election results are coarse signal that don't (directly) provide information about which issues matter

Beyond Politics...

There are other issues where surveys might be our only option...

- Public health monitoring (where to target interventions?)
- Poverty rates (could use IRS data, but miss people who don't file)
- Consumer surveys (which groups are likely to use Widgetly?)

Potential Solutions

1. Use administrative data that covers the full population
2. Combine multiple surveys together
3. Use a statistical model to generate an estimate

Combining Surveys

- We could use Roper Center iPoll (or other resources) to find many surveys that ask the same/similar question
- Combine the datasets together
- The hope is we have enough total sample size to get reasonable estimates for subgroups of interest

Combining Surveys

Advantages?

Combining Surveys

Advantages?

- Conceptually simple
- Doesn't require modeling assumptions
- Reduces impact of outlier surveys (think: poll averaging)

Combining Surveys

Advantages?

- Conceptually simple
- Doesn't require modeling assumptions
- Reduces impact of outlier surveys (think: poll averaging)

Disadvantages?

Combining Surveys

Advantages?

- Conceptually simple
- Doesn't require modeling assumptions
- Reduces impact of outlier surveys (think: poll averaging)

Disadvantages?

- May have to combine surveys over a long time span
  - Limits ability to see change over time
  - Can't investigate questions that aren't commonly included on surveys
- Often hard to find raw data
- Still might not have enough sample for small subgroups

Combining Surveys

Advantages?

- Conceptually simple
- Doesn't require modeling assumptions
- Reduces impact of outlier surveys (think: poll averaging)

Disadvantages?

- May have to combine surveys over a long time span
  - Limits ability to see change over time
  - Can't investigate questions that aren't commonly included on surveys
- Often hard to find raw data
- Still might not have enough sample for small subgroups

Takeaway: if feasible, not a bad solution. But often infeasible.

Potential Solutions

1. Use administrative data that covers the full population
2. Combine multiple surveys together
3. **Use a statistical model to generate an estimate**

Statistical Models for Small Subgroup Estimation

- In our survey, we may not have many (or any!) people in the subgroup we're interested in
- But maybe we have a large enough survey that we can fit a statistical model to *predict* the value as a function of demographic variables

Statistical Models for Small Subgroup Estimation

General framework: **multilevel regression** and **poststratification** (a.k.a. MRP)

Statistical Models for Small Subgroup Estimation

General framework: **multilevel regression** and **poststratification** (a.k.a. MRP)

Step 1. Estimate a **regression** model for the outcome variable as a function of demographic and geographic variables

Statistical Models for Small Subgroup Estimation

General framework: **multilevel regression** and **poststratification** (a.k.a. MRP)

Step 1. Estimate a **regression** model for the outcome variable as a function of demographic and geographic variables

Step 2. Use the model to *predict* outcome for every combination of demographic variables

Statistical Models for Small Subgroup Estimation

General framework: **multilevel regression** and **poststratification** (a.k.a. MRP)

Step 1. Estimate a **regression** model for the outcome variable as a function of demographic and geographic variables

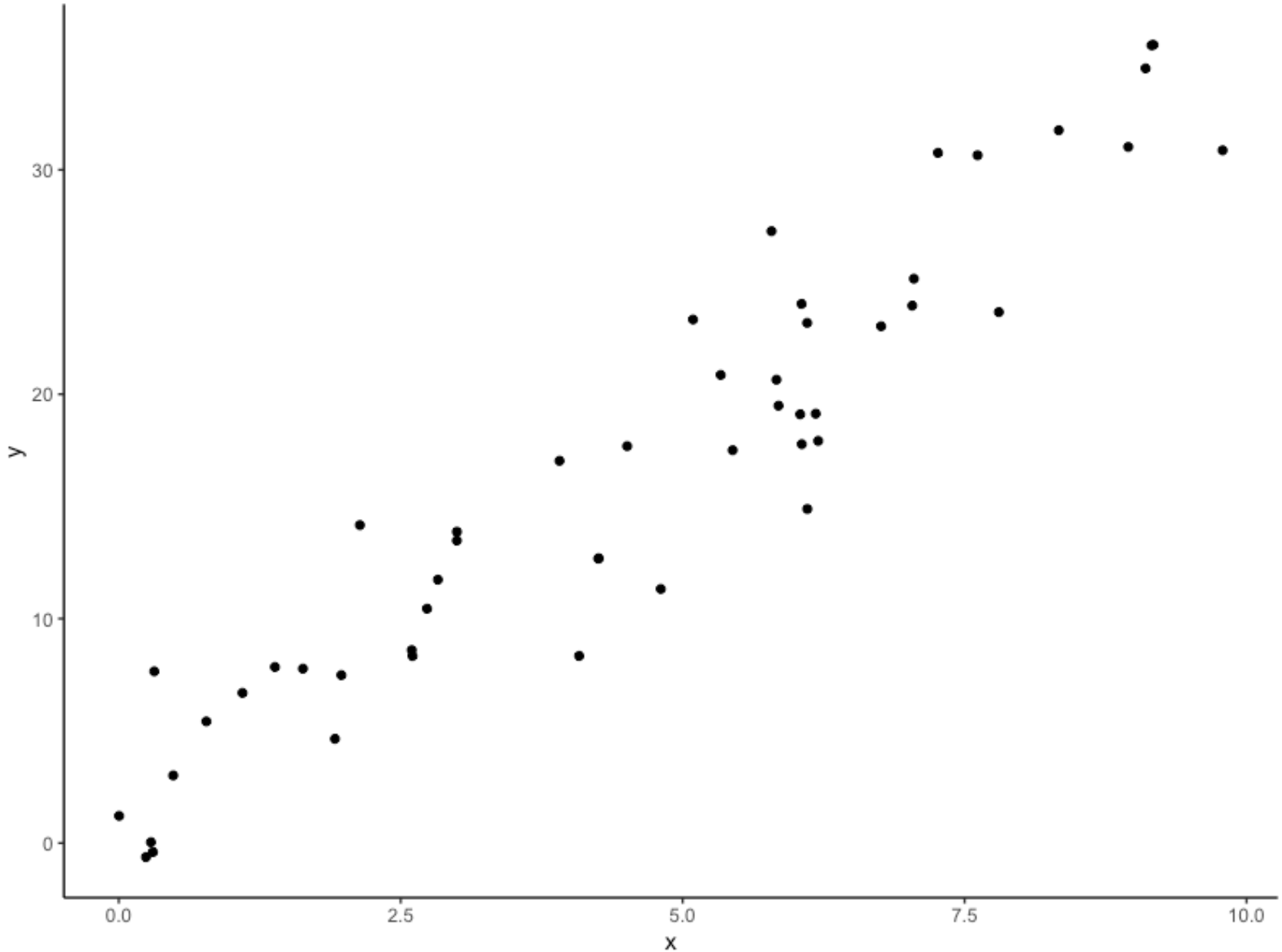
Step 2. Use the model to *predict* outcome for every combination of demographic variables

Step 3. Take weighted mean of the cells, using weights derived from the Census (**poststratification**)

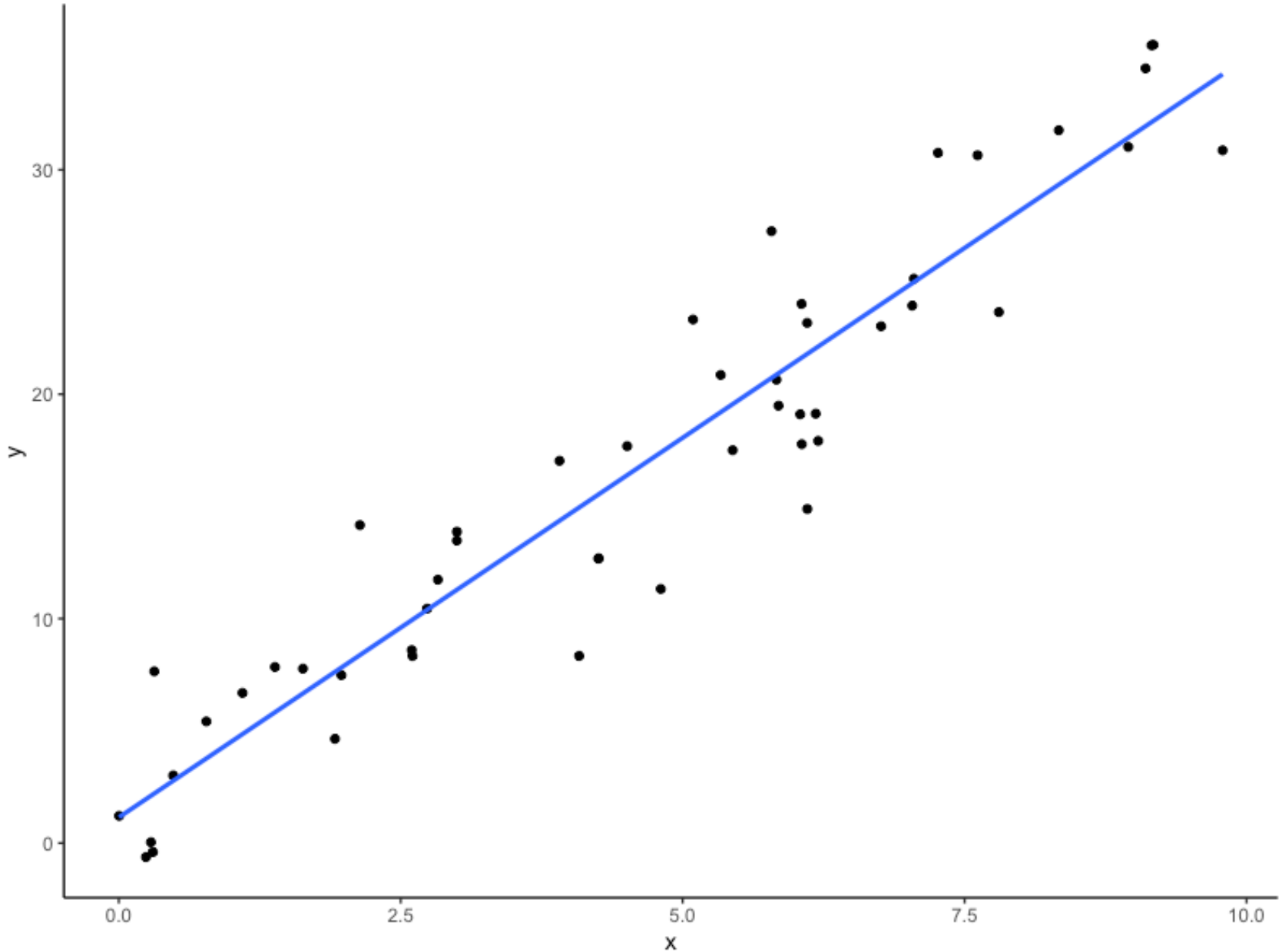
MRP: Intuition

- Demographic and geographic variables are often highly predictive of attitudes and behavior
- We might not have many Black women in Georgia in our survey, but they are likely to share some similarities with other similar groups: Black women nationwide, non-Black women in Georgia, all Black voters nationwide
- If 90% of Black voters nationwide will vote for the Democratic Senate candidate, that gives us (imperfect!) information about how Black women in Georgia will vote
- We can build this intuition into a statistical model, letting the data tell us how "similar" different groups are

Regression: Refresher



Regression: Refresher



Regression: Refresher

Regression finds a function that gives a "good" prediction of the dependent variable y, as a function of "independent variables" $x_1, x_2, \ldots$

Regression: Refresher

Regression finds a function that gives a "good" prediction of the dependent variable y, as a function of "independent variables" $x_1, x_2, \ldots$

Prediction: Given estimated regression coefficients $(\hat{b}_0, \hat{b}_1, \hat{b}_,2, \hat{b}_3, \ldots)$, we can *predict* what y would be for *any* combination of the independent variables, even if that combination isn't observed in the data

Step 1: Build A Regression Model Predicting the Outcome

- Set up a regression model to predict the outcome of interest
- Should include variables that are highly predictive and which you have population-level data on (e.g. from Census)
- Standard variables include age, race, gender, state, education, etc.
- Can also include aggregate geographic variables (e.g. state-level vote share or demographic data)

A Simple Regression Predicting Senate Vote

```
Call:
lm(formula = dem_twoparty_sen ~ gender + agegrp + race + educ,
    data = sm, weights = wt_4)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-1.53977 -0.37234  0.06586  0.33209  1.38199

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.753629   0.011429  65.938  < 2e-16 ***
gendermale       -0.145225   0.003881 -37.418  < 2e-16 ***
agegrp30-39      -0.051410   0.007810  -6.583 4.66e-11 ***
agegrp40-49      -0.083954   0.007658 -10.963  < 2e-16 ***
agegrp50-64      -0.095186   0.006872 -13.851  < 2e-16 ***
agegrp65-74      -0.026459   0.007645  -3.461 0.000539 ***
agegrp75+        -0.041998   0.008123  -5.170 2.34e-07 ***
raceblack         0.260466   0.011407  22.834  < 2e-16 ***
racehispanic      0.019125   0.011142   1.717 0.086073 .
raceother        -0.118912   0.015078  -7.886 3.17e-15 ***
racewhite        -0.121122   0.009853 -12.293  < 2e-16 ***
educHS or less   -0.127961   0.005544 -23.082  < 2e-16 ***
educpostgrad      0.109319   0.006465  16.910  < 2e-16 ***
educsome college -0.075101   0.005210 -14.416  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4226 on 60656 degrees of freedom
  (127371 observations deleted due to missingness)
Multiple R-squared:  0.097,     Adjusted R-squared:  0.09681
F-statistic: 501.2 on 13 and 60656 DF,  p-value: < 2.2e-16
```

Step 2: Predict Outcomes for Demographic Cells

- We'll use Census data to get a table of the number of people in each cell used in the regression model
- E.g. the Census data should have the number of people in each combination of age, race, gender, education, etc.
- Predict the average outcome variable for each of these groups using the model

Predicting Vote Senate Vote

	gender	agegrp	race	educ	est_prop
1	male	75+	white	some college	0.0072733425
2	male	75+	white	HS or less	0.0132364021
3	male	40-49	asian	college	0.0015663066
4	female	75+	other	HS or less	0.0004777821
5	male	75+	hispanic	HS or less	0.0021644024
6	female	40-49	other	some college	0.0008785699
7	male	18-29	asian	college	0.0015592609
8	female	75+	hispanic	college	0.0002295887
9	male	18-29	hispanic	college	0.0017161588
10	male	40-49	hispanic	college	0.0015982766
11	male	40-49	black	HS or less	0.0043840520
12	female	50-64	black	HS or less	0.0064364139
13	male	40-49	hispanic	postgrad	0.0008477186
14	female	30-39	other	some college	0.0010773810
15	female	65-74	other	postgrad	0.0001698397
16	female	18-29	other	postgrad	0.0001572979
17	male	50-64	asian	HS or less	0.0018890863
18	male	30-39	other	college	0.0006465848
19	female	40-49	black	college	0.0017391215
20	male	75+	black	college	0.0002162934

Predicting Vote Senate Vote

```
ps.simple$senate_dem = predict(mod, ps.simple)
```

	gender	agegrp	race	educ	est_prop	senate_dem
1	male	75+	white	some college	0.0072733425	0.3701831
2	male	75+	white	HS or less	0.0132364021	0.3173229
3	male	40-49	asian	college	0.0015663066	0.5244495
4	female	75+	other	HS or less	0.0004777821	0.4647572
5	male	75+	hispanic	HS or less	0.0021644024	0.4575700
6	female	40-49	other	some college	0.0008785699	0.4756610
7	male	18-29	asian	college	0.0015592609	0.6084037
8	female	75+	hispanic	college	0.0002295887	0.7307564
9	male	18-29	hispanic	college	0.0017161588	0.6275292
10	male	40-49	hispanic	college	0.0015982766	0.5435749
11	male	40-49	black	HS or less	0.0043840520	0.6569538
12	female	50-64	black	HS or less	0.0064364139	0.7909472
13	male	40-49	hispanic	postgrad	0.0008477186	0.6528943
14	female	30-39	other	some college	0.0010773810	0.5082054
15	female	65-74	other	postgrad	0.0001698397	0.7175771
16	female	18-29	other	postgrad	0.0001572979	0.7440357
17	male	50-64	asian	HS or less	0.0018890863	0.3852564
18	male	30-39	other	college	0.0006465848	0.4380815
19	female	40-49	black	college	0.0017391215	0.9301402
20	male	75+	black	college	0.0002162934	0.8268716

Step 3: Take Weighted Mean of the Cells

- Finally, we use the population-level proportions to aggregate up to the subgroup we're interested in
- E.g., if we care about all college-educated voters, we average together all categories that include college-educated voters
- Use a simple weighted mean function to do this
- Can only aggregate to variables that are included in the model!

Aggregating Up

```
> black_women_dem_share = ps.simple %>%
+   filter(race == "black", gender == "female") %>%
+   summarise(estimate = weighted.mean(senate_dem, w = est_prop))
> black_women_dem_share
# A tibble: 1 × 1
  estimate
     <dbl>
1    0.891
```

Considerations

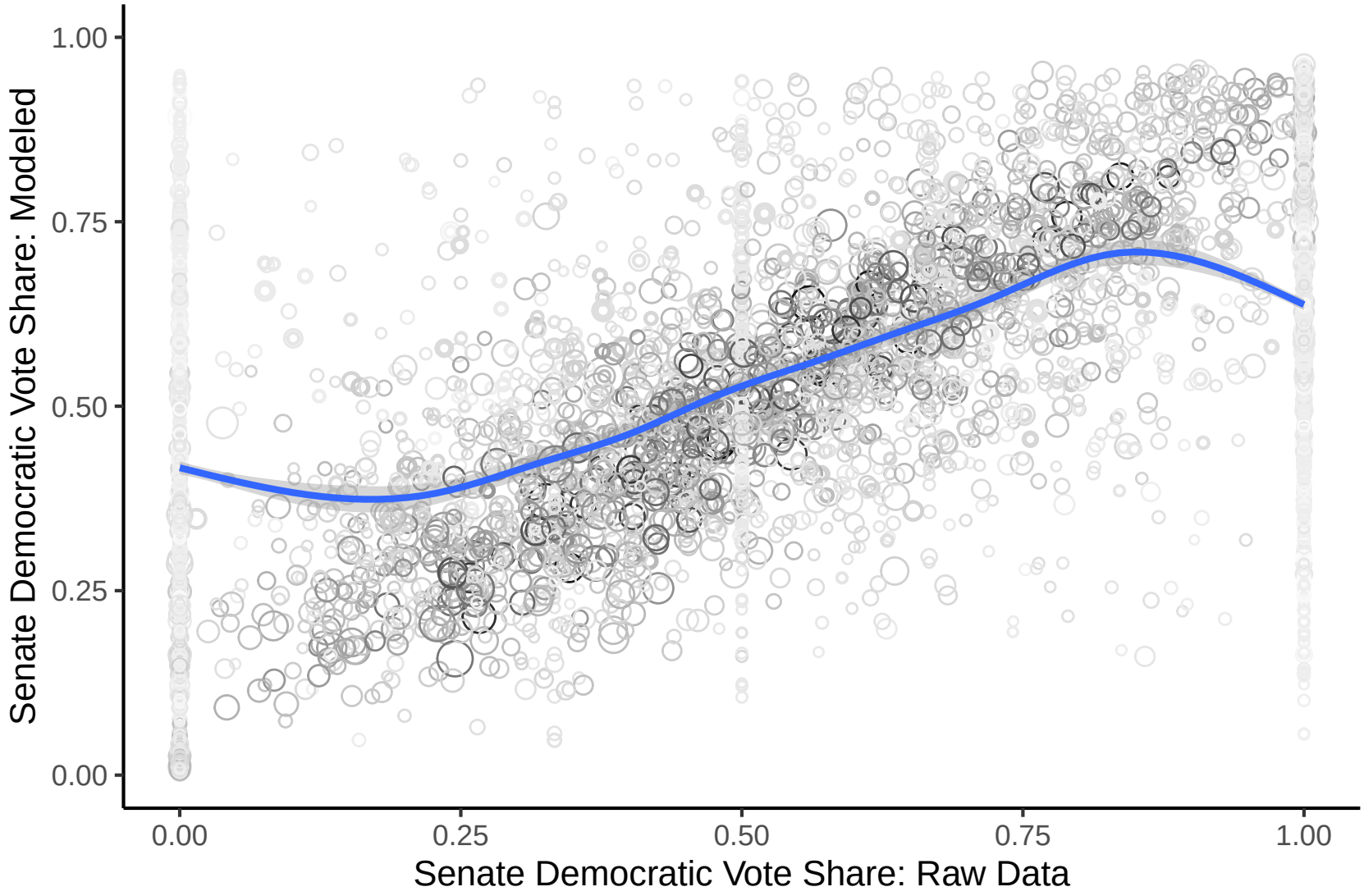
- Need a strongly predictive model: if there is lots of heterogeneity within groups, won't work as well
- General would want to estimate more flexible model to account for more variables
- May want to include geographic-level predictors, such as past election results
- Usually use some kind of *regularization*: multilevel model or machine learning methods
- Assumes relationships between variables are the same across states

Full Model on PORES/SurveyMonkey Data

Estimate Dem. vote share as a function of:

- age, race, gender, education, state
- race $\times$ education
- white/nonwhite $\times$ education $\times$ age
- State-level variables: % college educated, % nonwhite, % Latino, 2020 Biden vote share, state median income

Total of $> 18,000$ cells.


Senate Democratic Vote Share: Modeled
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Senate Democratic Vote Share: Raw Data

Black Women in Georgia Senate Race

Raw data ($N = 411$): 94.4% of Black women in Georgia will vote for Warnock.
MRP estimate: 87.7%

Application: Simonovits and Payson (2023)

Big-picture question: **Does local control of policy lead to policy outcomes closer to citizens' preferences?**

Application: Simonovits and Payson (2023)

Big-picture question: **Does local control of policy lead to policy outcomes closer to citizens' preferences?**

Background:

- "Voting with your feet"
- Decentralization $\rightsquigarrow$ tight congruence between (local) public preferences and policy outcomes
- But costly to measure public opinion in small geographies

Setting: Minimum wage

- Federal minimum wage stagnant at \$7.25 since 2009 (equivalent to \$10.50 today but not indexed to inflation)
- States and cities have passed minimum wage laws, but ability of cities to set minimum wage varies substantiall
- 25 states ban cities from setting local minimum wage
- Question: How similar are minimum wages to public preferences in cities that do and do not set their own policy?

Necessary Data

Necessary Data

State and local minimum wage laws

- sourced from UC Berkeley's Labor Center, Economic Policy Institute, National Conference of State Legislatures

Necessary Data

State and local minimum wage laws

- sourced from UC Berkeley's Labor Center, Economic Policy Institute, National Conference of State Legislatures

Public preferences for minimum wage via survey

- "In your view, what should be the minimum hourly wage in the town or city you currently live in?"
- Large online survey, $N \approx 18,000$
- Still, there are a lot of cities: over 75% of cities have only one respondent

MRP Approach

Use MRP model to estimate local public opinion:

- fit regression model predicting opinion as a function of gender, education, race, ethnicity, age, income, and geographic predictors (median income, rent, etc.)
- poststratify using Census data

Results: Distribution of Preferences

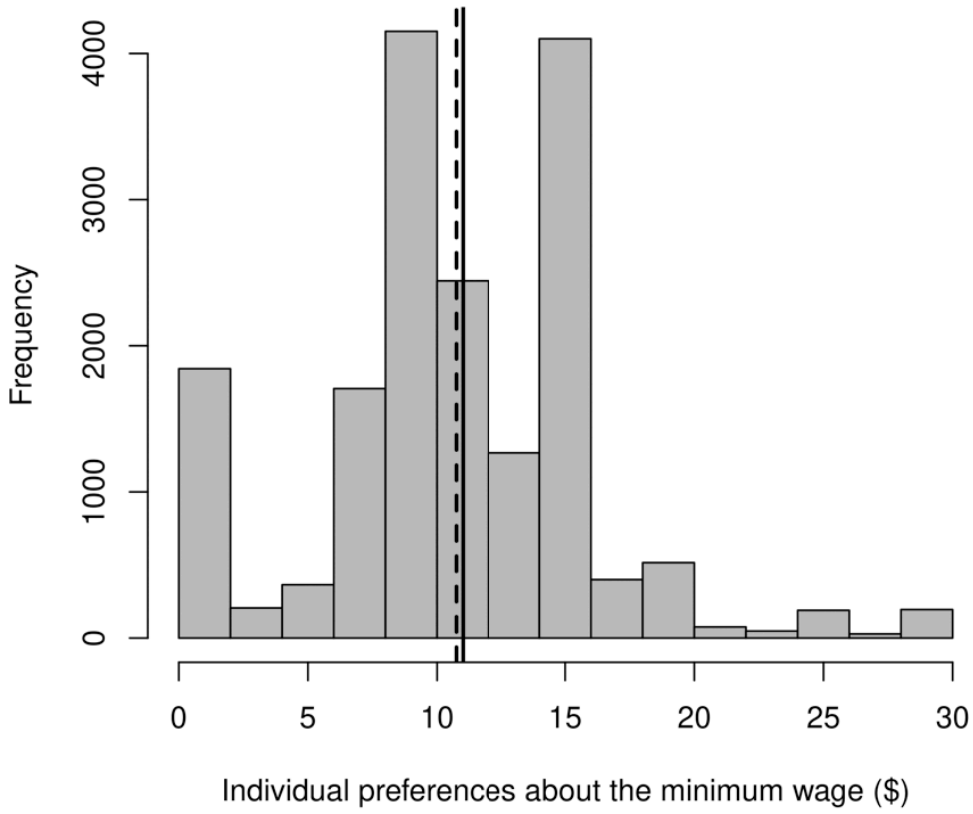


Figure 1: Distribution of individual preferences for local minimum wages.

Note: The mean is marked by the solid line and the median by the dashed line.

Results: Validation Against Ballot Measure in FL

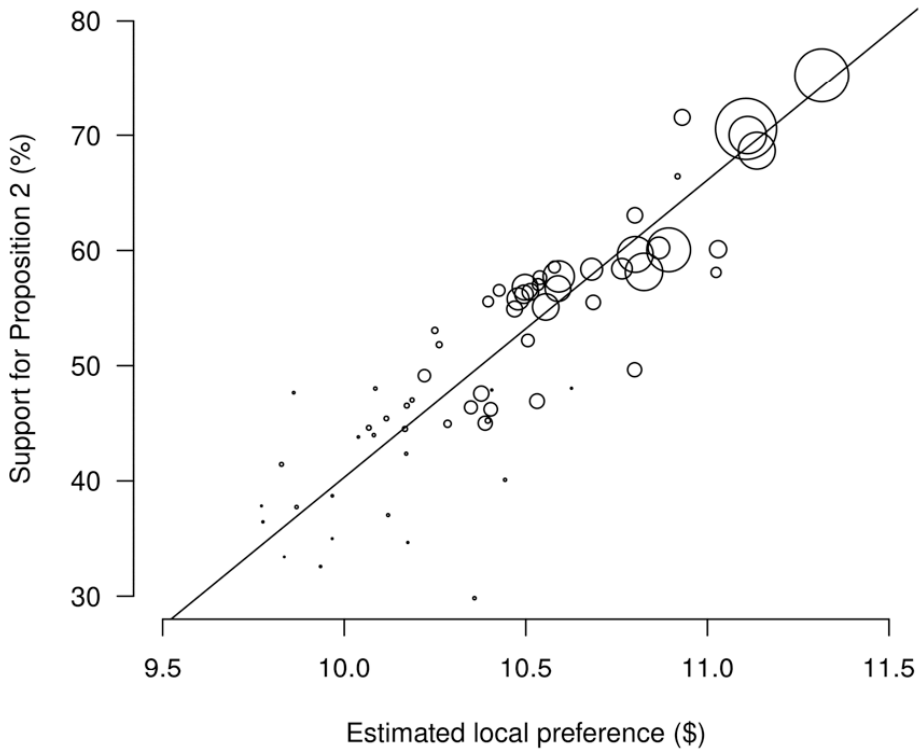


Figure 2: Validating MrsP estimates with Florida election data.

Note: Circles plot county-level support for Proposition 2 in Florida against estimated city minimum wage preferences aggregated to the county level. Circles are proportional to the square-root of population sizes. Line shows best fitting regression estimated via WLS.

Results: Sample of Cities

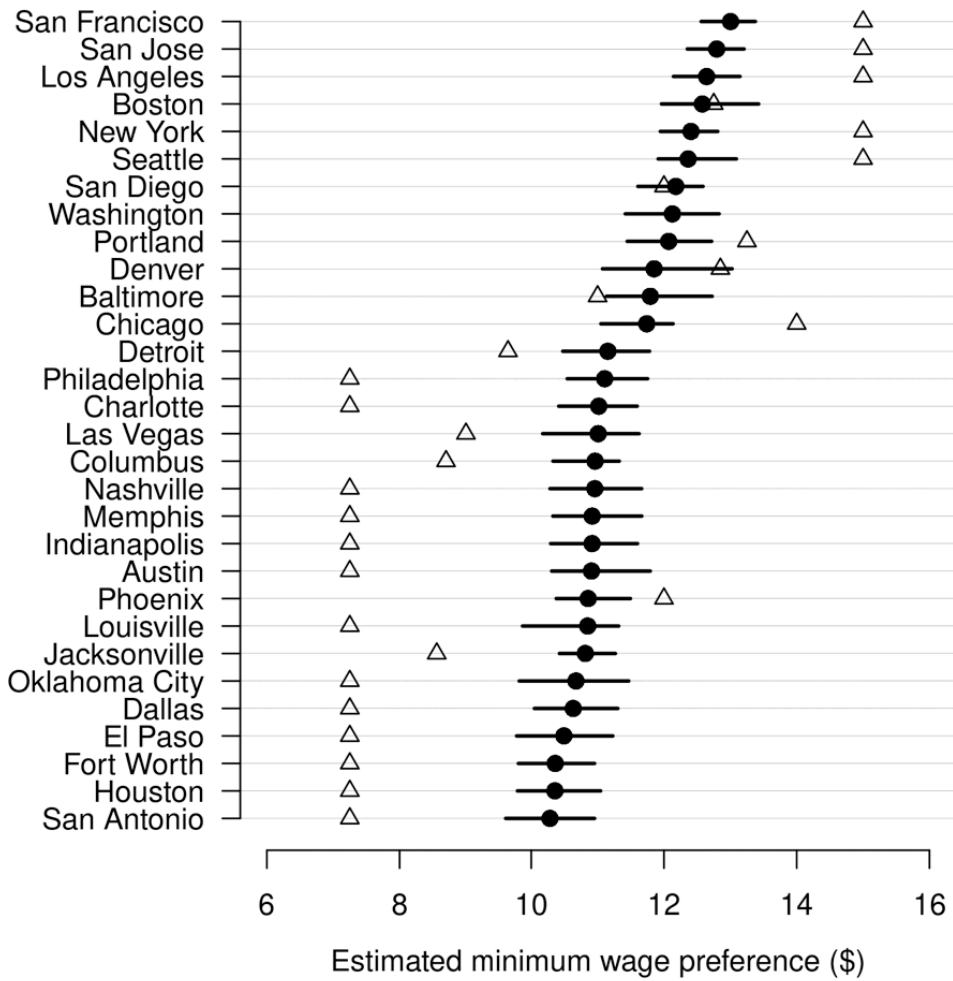


Figure 3: Estimates of city minimum wage preferences.

Note: Solid dots denote our city-level preference estimates with 95% confidence intervals. Triangles represent effective minimum wages.

Results: All Cities

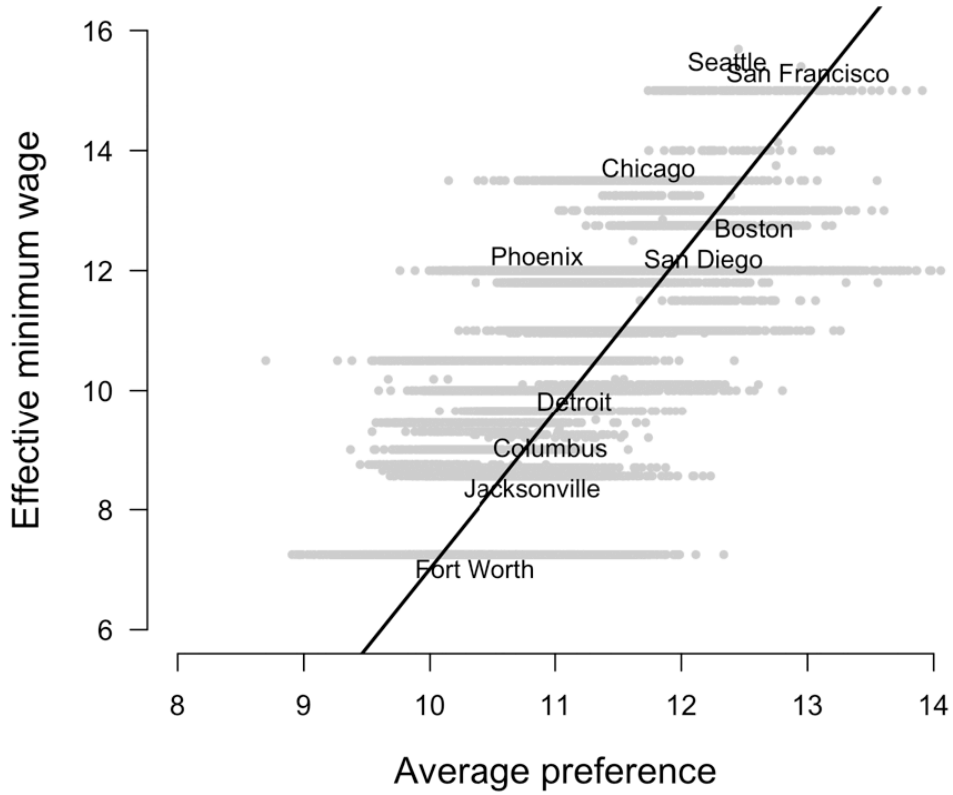


Figure 4: Minimum wage preferences and local policies.

Note: Shows estimated city minimum wage preferences plotted against effective minimum wages with a regression line estimated via WLS.

Results: By Local Control

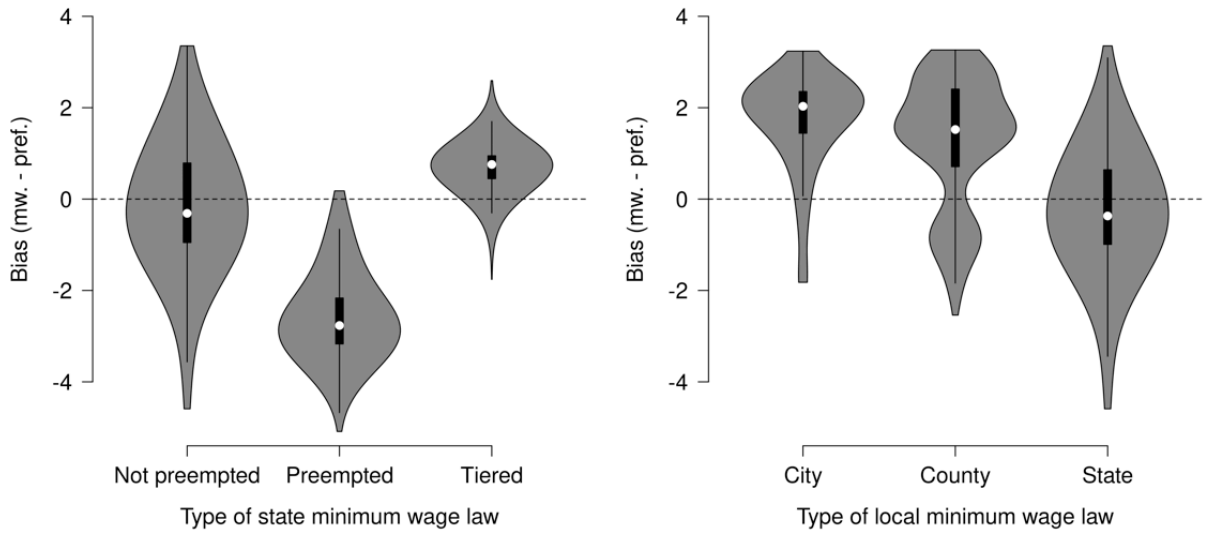


Figure 5: Policy bias and decentralization.

Note: Violin plots in the left panel visualize the distribution of policy bias (i.e., policy — average preference) in localities across states with different state laws. The right panel shows the distribution of policy bias across cities within states with no preemption based on the type of minimum wage law adopted.

Discussion

- What assumptions do we need to make in order to believe the estimates of public opinion?
- What do these results suggest for policymakers and activists who want to increase minimum wages?
- How generalizable do you think these results are across policy areas?