Demographics and Turnout

GOV 1347 Lab: Week V

Matthew E. Dardet

Harvard University

October 2, 2024

Check-In

- Any questions about the course so far?

Check-In

- Any questions about the course so far?
- VP Debate: any assessments/thoughts/prognostications?

This Week's Goal

Main Question: What are the connections, if any, between demographics and voter turnout? Can demographics and turnout be used to predict election outcomes? If so, how and how well?

1. **Turnout & Demographics**
2. **Trees, Forests, Oh My! Replicating Kim & Zilinsky (2023).**
3. **Voter File: Forecasting Meets Big Data**
4. **Simulations**

Section 1

Turnout & Demographics

Who Votes?

Any guesses?

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.
     - Find that **education** is the key demographic variable influencing turnout. Why?

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.
     - Find that **education** is the key demographic variable influencing turnout. Why? Information, skills, civic virtue.

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.
     - Find that **education** is the key demographic variable influencing turnout. Why? Information, skills, civic virtue.
     - Also find that **age**, **voter registration laws**, **marriage**, and **occuption** matter.

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.
     - Find that **education** is the key demographic variable influencing turnout. Why? Information, skills, civic virtue.
     - Also find that **age**, **voter registration laws**, **marriage**, and **occuption** matter.
  2. Rosenstone & Hansen (1993)
     - Use data from the American National Eleciton Studies (ANES).

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.
     - Find that **education** is the key demographic variable influencing turnout. Why? Information, skills, civic virtue.
     - Also find that **age**, **voter registration laws**, **marriage**, and **occupation** matter.
  2. Rosenstone & Hansen (1993)
     - Use data from the American National Eleciton Studies (ANES).
     - Find that **mobilization** efforts through social networks are key factors determining voter turnout.

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.
     - Find that **education** is the key demographic variable influencing turnout. Why? Information, skills, civic virtue.
     - Also find that **age**, **voter registration laws**, **marriage**, and **occuption** matter.
  2. Rosenstone & Hansen (1993)
     - Use data from the American National Eleciton Studies (ANES).
     - Find that **mobilization** efforts through social networks are key factors determining voter turnout.
     - White, wealthy, educated people participate the most in voting and politics generally.

Who Votes?

Any guesses?

- Early points of view:
  1. Wolfinger & Rosenstone (1980)
     - Use Census data from 1972 and 1974 and run simple OLS regressions.
     - Find that **education** is the key demographic variable influencing turnout. Why? Information, skills, civic virtue.
     - Also find that **age**, **voter registration laws**, **marriage**, and **occuption** matter.
  2. Rosenstone & Hansen (1993)
     - Use data from the American National Eleciton Studies (ANES).
     - Find that **mobilization** efforts through social networks are key factors determining voter turnout.
     - White, wealthy, educated people participate the most in voting and politics generally.
     - Voter turnout declined in the 1960s and 1980s, possibly due to lack of mobilization efforts, impersonal and candidate-centered politics.

More Recent Analyses of Demographics, Turnout, and Voting

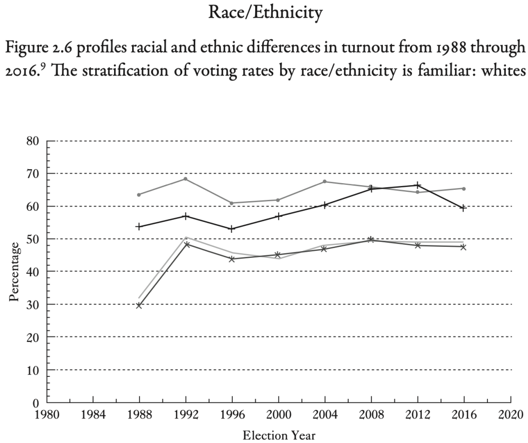
These older studies found that demographics matter for voter turnout.

More Recent Analyses of Demographics, Turnout, and Voting

These older studies found that demographics matter for voter turnout. What do Shaw & Petrocik (2020) find?

More Recent Analyses of Demographics, Turnout, and Voting

These older studies found that demographics matter for voter turnout. What do Shaw & Petrocik (2020) find?



Race/Ethnicity

Figure 2.6 profiles racial and ethnic differences in turnout from 1988 through 2016.[9] The stratification of voting rates by race/ethnicity is familiar: whites

Turnout by Education

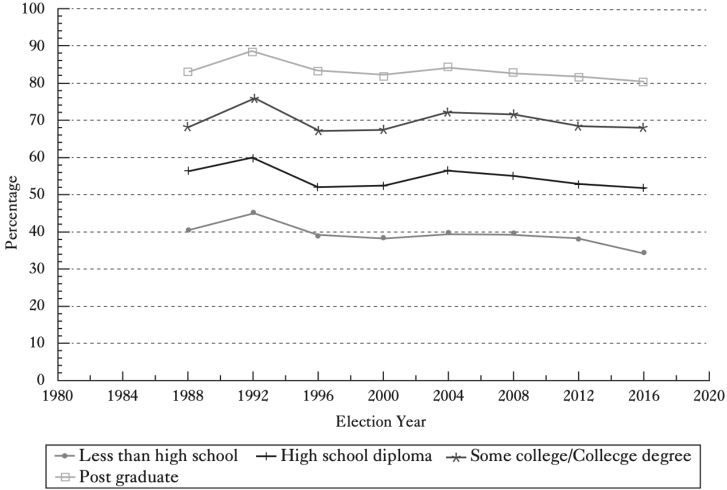


FIGURE 2.7 Turnout by education level, 1988–2016.

Turnout by Age

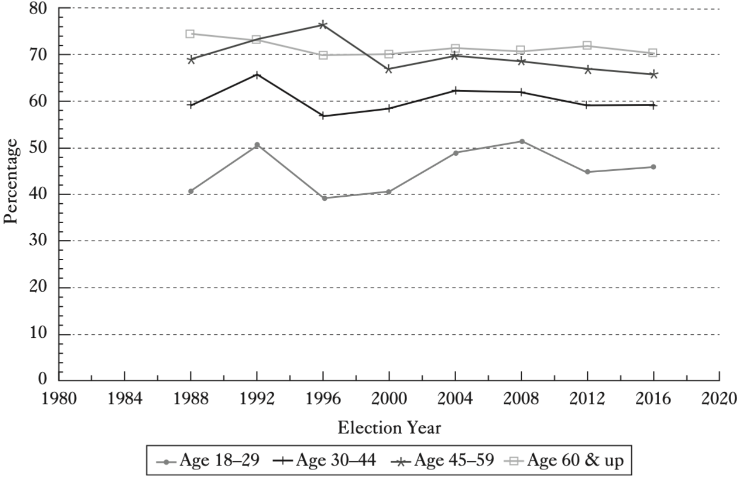


FIGURE 2.8 Turnout by age, 1988–2016.

What About Turnout and Vote *Choice*?

- **The Bias Thesis**:

What About Turnout and Vote *Choice*?

- **The Bias Thesis**:
    - Increase in turnout increases vote share for Democrats.

What About Turnout and Vote *Choice*?

- **The Bias Thesis**:
  - Increase in turnout increases vote share for Democrats.
  - Decrease in turnout decreases vote share for Democrats.

What About Turnout and Vote *Choice*?

- **The Bias Thesis**:
  - Increase in turnout increases vote share for Democrats.
  - Decrease in turnout decreases vote share for Democrats.
- Is it real?

What About Turnout and Vote *Choice*?

- **The Bias Thesis**:
  - Increase in turnout increases vote share for Democrats.
  - Decrease in turnout decreases vote share for Democrats.
- Is it real?

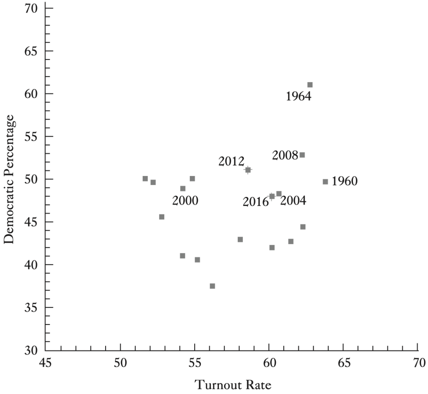
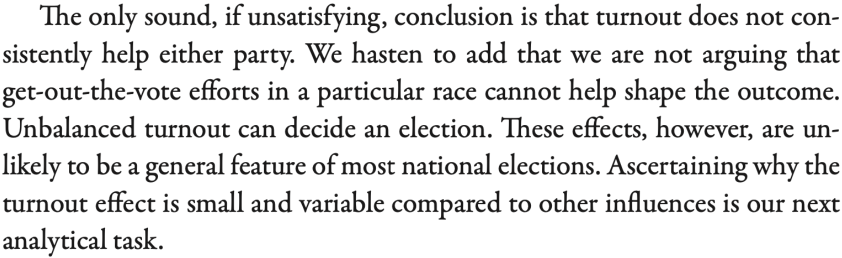


FIGURE 4.1 Turnout and the presidential vote, 1948–2016.

What About Turnout and Vote *Choice*?

- **The Bias Thesis**:
    - Increase in turnout increases vote share for Democrats.
    - Decrease in turnout decreases vote share for Democrats.
- Is it real?



The only sound, if unsatisfying, conclusion is that turnout does not consistently help either party. We hasten to add that we are not arguing that get-out-the-vote efforts in a particular race cannot help shape the outcome. Unbalanced turnout can decide an election. These effects, however, are unlikely to be a general feature of most national elections. Ascertaining why the turnout effect is small and variable compared to other influences is our next analytical task.

What About Demographics and Vote *Choice*?

- What do Kim & Zilinsky (2023) say about demographics and vote choice?

What About Demographics and Vote *Choice*?

- What do Kim & Zilinsky (2023) say about demographics and vote choice?
  - Logistic regressions and random forest models with training data encompassing only demographics from the ANES between 1952-2020.

What About Demographics and Vote *Choice*?

- What do Kim & Zilinsky (2023) say about demographics and vote choice?
    - Logistic regressions and random forest models with training data encompassing only demographics from the ANES between 1952-2020.
    - Models predict 63.9% of two-party vote choices and 63.4% of partisan IDs correctly out-of-sample.

What About Demographics and Vote *Choice*?

- What do Kim & Zilinsky (2023) say about demographics and vote choice?
    - Logistic regressions and random forest models with training data encompassing only demographics from the ANES between 1952-2020.
    - Models predict 63.9% of two-party vote choices and 63.4% of partisan IDs correctly out-of-sample.
    - They conclude that demographics "reveal little about voting and partisanship."

Section 2

Kim & Zilinsky (2023) Replication and Analysis

Replication: Setup

- Demographic Variables (Features):
  - *Basic*: Age, Gender, Race, Income, Education
  - *Extension*: Urbanicity, Region, Southern, Employmet, Religion, Homeownership, Marriage

Replication: Setup

- Demographic Variables (Features):
  - *Basic*: Age, Gender, Race, Income, Education
  - *Extension*: Urbanicity, Region, Southern, Employmet, Religion, Homeownership, Marriage
- Outcome Variables (Response):
  - Partisanship (Democrat/Republican)
  - Vote Choice (Democrat/Republican)

Replication: Setup

- Demographic Variables (Features):
  - *Basic*: Age, Gender, Race, Income, Education
  - *Extension*: Urbanicity, Region, Southern, Employmet, Religion, Homeownership, Marriage
- Outcome Variables (Response):
  - Partisanship (Democrat/Republican)
  - Vote Choice (Democrat/Republican)
- Going to compare predictability using OLS, logistic regression, and random forest models.

Logistic Regression: Motivation

- Let's say we want to predict a binary outcome like turnout or whether someone votes for a Republican candidate.

Logistic Regression: Motivation

- Let's say we want to predict a binary outcome like turnout or whether someone votes for a Republican candidate.
- When we fit a linear regression with OLS, $Y = \beta X$, there are no restrictions on Y.

Logistic Regression: Motivation

- Let's say we want to predict a binary outcome like turnout or whether someone votes for a Republican candidate.
- When we fit a linear regression with OLS, $Y = \beta X$, there are no restrictions on Y. What could go wrong with that?

Logistic Regression: Motivation

- Let's say we want to predict a binary outcome like turnout or whether someone votes for a Republican candidate.
- When we fit a linear regression with OLS, $Y = \beta X$, there are no restrictions on Y. What could go wrong with that?
- In these cases, Y could theoretically have prediction intervals $(-\infty, \infty)$ beyond the bounds of the variables support.

Logistic Regression: Motivation

- Let's say we want to predict a binary outcome like turnout or whether someone votes for a Republican candidate.
- When we fit a linear regression with OLS, $Y = \beta X$, there are no restrictions on Y. What could go wrong with that?
- In these cases, Y could theoretically have prediction intervals $(-\infty, \infty)$ beyond the bounds of the variables support.
- In practice, this can occur when we are extrapolating, but also when there is sparse data (e.g., when we fit linear regression model on state-level polls).

Logistic Regression: Overview

- Return to our general prediction framework $Y = f(X)$.

Logistic Regression: Overview

- Return to our general prediction framework $Y = f(X)$.
- With binomial logistic regression, we use a **link function** f, known as the inverse logistic function a.k.a. the log odds function, to bound $(-\infty, \infty)$ to $(0, 1)$.

Logistic Regression: Overview

- Return to our general prediction framework $Y = f(X)$.
- With binomial logistic regression, we use a **link function** f, known as the inverse logistic function a.k.a. the log odds function, to bound $(-\infty, \infty)$ to $(0, 1)$.

$$\begin{aligned} f(x) &= \text{logit}^{-1}(x) \\ &= \text{logistic}(x) \end{aligned}$$

Logistic Regression: Overview

- Return to our general prediction framework $Y = f(X)$.
- With binomial logistic regression, we use a **link function** f, known as the inverse logistic function a.k.a. the log odds function, to bound $(-\infty, \infty)$ to $(0, 1)$.

$$\begin{aligned} f(x) &= \text{logit}^{-1}(x) \\ &= \text{logistic}(x) \\ &= \frac{1}{1 + \exp(-x)} \end{aligned}$$

Logistic Regression: Overview

- Return to our general prediction framework $Y = f(X)$.
- With binomial logistic regression, we use a **link function** f, known as the inverse logistic function a.k.a. the log odds function, to bound $(-\infty, \infty)$ to $(0, 1)$.

$$\begin{aligned} f(x) &= \text{logit}^{-1}(x) \\ &= \text{logistic}(x) \\ &= \frac{1}{1+\exp(-x)} \\ &= \frac{\exp(x)}{\exp(x)+1} \quad . \end{aligned}$$

Logistic Regression: Overview

- Return to our general prediction framework $Y = f(X)$.
- With binomial logistic regression, we use a **link function** f, known as the inverse logistic function a.k.a. the log odds function, to bound $(-\infty, \infty)$ to $(0, 1)$.

$$
\begin{aligned}
f(x) &= \text{logit}^{-1}(x) \\
&= \text{logistic}(x) \\
&= \frac{1}{1 + \exp(-x)} \\
&= \frac{\exp(x)}{\exp(x) + 1} \quad .
\end{aligned}
$$

Fun fact: this function is also known as the **sigmoid** $\sigma(x)$, and is used in **neural networks**.

Logistic Regression: Mechanics

- In the univariate case, our data $X = \beta_0 + \beta_1 X$.
- The logistic regression model is then:

$$Y = \mathrm{logit}^{-1}(\beta_0 + \beta_1 X)$$

Logistic Regression: Mechanics

- In the univariate case, our data $X = \beta_0 + \beta_1 X$.
- The logistic regression model is then:

$$\begin{aligned} Y &= \text{logit}^{-1}(\beta_0 + \beta_1 X) \\ &= \frac{\exp(\beta_0 + \beta_1 X)}{\exp(\beta_0 + \beta_1 X) + 1} \quad . \end{aligned}$$

Logistic Regression: Mechanics

- In the univariate case, our data $X = \beta_0 + \beta_1 X$.
- The logistic regression model is then:

$$\begin{aligned} Y &= \text{logit}^{-1}(\beta_0 + \beta_1 X) \\ &= \frac{\exp(\beta_0 + \beta_1 X)}{\exp(\beta_0 + \beta_1 X) + 1} \quad . \end{aligned}$$

- Coefficients interpreted in terms of log odds ratios and can map onto predicted probabilities (less important for classification ↔ binary prediction).

Logistic Regression: Mechanics

- In the univariate case, our data $X = \beta_0 + \beta_1 X$.
- The logistic regression model is then:

$$Y = \text{logit}^{-1}(\beta_0 + \beta_1 X) = \frac{\exp(\beta_0 + \beta_1 X)}{\exp(\beta_0 + \beta_1 X) + 1} \quad .$$

- Coefficients interpreted in terms of log odds ratios and can map onto predicted probabilities (less important for classification ↔ binary prediction).
- Can be extended to multiple outcome categories — multinomial logistic regression.

Logistic Regression: Implementation

- In `R`, we can use the `glm()` function with `family = "binomial"` to fit binomial logistic regression models.

Logistic Regression: Implementation

- In `R`, we can use the `glm()` function with `family = "binomial"` to fit binomial logistic regression models.
- `Live coding!`

Random Forests: Motivation

- But first, consider the humble tree

Random Forests: Motivation

- But first, consider the humble tree

Decision Tree: Overview

- The humble classification and regression tree (CART) that is.

Decision Tree: Overview

- The humble classification and regression tree (CART) that is.
- Tree-based methods work by taking **feature space**——set of all your independent variables——and **partioning** the feature space into regions wiht similar response values by learning a set of **splitting rules**.

Decision Tree: Overview

- The humble classification and regression tree (CART) that is.
- Tree-based methods work by taking **feature space**——set of all your independent variables——and **partioning** the feature space into regions wiht similar response values by learning a set of **splitting rules**.
- Determining optimal partition is what's known as an *NP-hard* problem.

Decision Tree: Overview

- The humble classification and regression tree (CART) that is.
- Tree-based methods work by taking **feature space**——set of all your independent variables——and **partioning** the feature space into regions wiht similar response values by learning a set of **splitting rules**.
- Determining optimal partition is what's known as an *NP-hard* problem.
- Can solve for all of these models using optimization.

Decision Tree: Overview

- The humble classification and regression tree (CART) that is.
- Tree-based methods work by taking **feature space**——set of all your independent variables——and **partioning** the feature space into regions wiht similar response values by learning a set of **splitting rules**.
- Determining optimal partition is what's known as an *NP-hard* problem.
- Can solve for all of these models using optimization.
    - In this case, want to minimize sum of squared error within each partition.

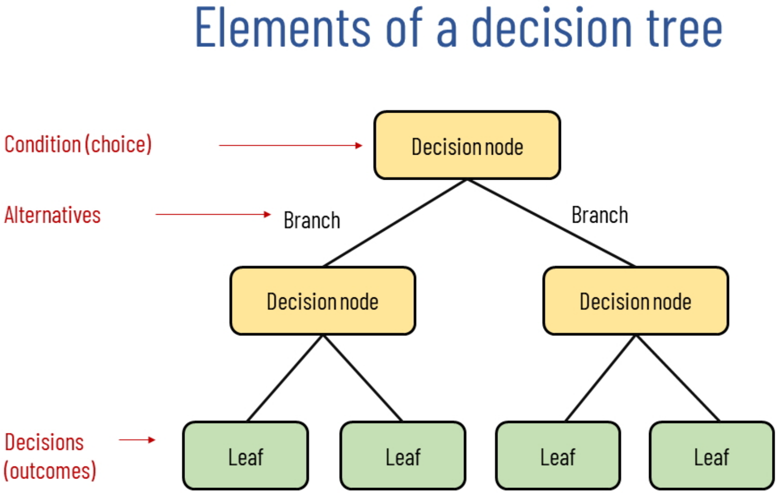
Decision Tree: Overview

- The humble classification and regression tree (CART) that is.
- Tree-based methods work by taking **feature space**——set of all your independent variables——and **partioning** the feature space into regions wiht similar response values by learning a set of **splitting rules**.
- Determining optimal partition is what's known as an *NP-hard* problem.
- Can solve for all of these models using optimization.
  - In this case, want to minimize sum of squared error within each partition.
  - Keep partitioning (binary) until a *stopping rule* is reached: number of observations per node.

Decision Tree: Overview

- The humble classification and regression tree (CART) that is.
- Tree-based methods work by taking **feature space**——set of all your independent variables——and **partioning** the feature space into regions wiht similar response values by learning a set of **splitting rules**.
- Determining optimal partition is what's known as an *NP-hard* problem.
- Can solve for all of these models using optimization.
    - In this case, want to minimize sum of squared error within each partition.
    - Keep partitioning (binary) until a *stopping rule* is reached: number of observations per node.
- Can get feature importance measure by determining reduction in loss function achieved by each variable at each split.

Decision Tree: Visualization



Elements of a decision tree

Condition (choice)

Alternatives

Decisions (outcomes)

Decision node

Branch

Branch

Decision node

Decision node

Leaf

Leaf

Leaf

Leaf

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!
- **Intuition:** have a bunch of unbiased classifiers, high variance. Averaging reduces variance but will be correlated. So, introduce additional sampling to induce independence.

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!
- **Intuition:** have a bunch of unbiased classifiers, high variance. Averaging reduces variance but will be correlated. So, introduce additional sampling to induce independence.
- **Idea:** independently reshuffle data m times, fit a bunch of trees, aggregate the results from individual trees → ensemble of trees → forest.

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!
- **Intuition:** have a bunch of unbiased classifiers, high variance. Averaging reduces variance but will be correlated. So, introduce additional sampling to induce independence.
- **Idea:** independently reshuffle data *m* times, fit a bunch of trees, aggregate the results from individual trees → ensemble of trees → forest.
  - Decorrelated trees reduce variance and increase predictive performance.

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!
- **Intuition:** have a bunch of unbiased classifiers, high variance. Averaging reduces variance but will be correlated. So, introduce additional sampling to induce independence.
- **Idea:** independently reshuffle data *m* times, fit a bunch of trees, aggregate the results from individual trees → ensemble of trees → forest.
    - Decorrelated trees reduce variance and increase predictive performance.
- **Terminology:** Bootstrapping and bagging:

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!
- **Intuition:** have a bunch of unbiased classifiers, high variance. Averaging reduces variance but will be correlated. So, introduce additional sampling to induce independence.
- **Idea:** independently reshuffle data *m* times, fit a bunch of trees, aggregate the results from individual trees → ensemble of trees → forest.
  - Decorrelated trees reduce variance and increase predictive performance.
- **Terminology:** Bootstrapping and bagging:
  - Bootstrapping is sampling with replacement (as opposed to *k*-fold cross validation).

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!
- **Intuition:** have a bunch of unbiased classifiers, high variance. Averaging reduces variance but will be correlated. So, introduce additional sampling to induce independence.
- **Idea:** independently reshuffle data *m* times, fit a bunch of trees, aggregate the results from individual trees → ensemble of trees → forest.
  - Decorrelated trees reduce variance and increase predictive performance.
- **Terminology:** Bootstrapping and bagging:
  - Bootstrapping is sampling with replacement (as opposed to *k*-fold cross validation).
  - Bagging is bootstrap aggregating: averaging the results of bootstrapped samples.

Random Forest: Motivation

- Same motivation as ensembling: One tree is great——many trees is better!
- **Intuition:** have a bunch of unbiased classifiers, high variance. Averaging reduces variance but will be correlated. So, introduce additional sampling to induce independence.
- **Idea:** independently reshuffle data *m* times, fit a bunch of trees, aggregate the results from individual trees → ensemble of trees → forest.
    - Decorrelated trees reduce variance and increase predictive performance.
- **Terminology:** Bootstrapping and bagging:
    - Bootstrapping is sampling with replacement (as opposed to *k*-fold cross validation).
    - Bagging is bootstrap aggregating: averaging the results of bootstrapped samples.
        - Random forests use an expanded version of bagging that limits search to random subsets of features to split on.

Random Forests: Implementation

- In `R`, we can use the `caret`, `mlr3`, `randomForest`, and/or `ranger` packages to fit random forest models.

Random Forests: Implementation

- In `R`, we can use the `caret`, `mlr3`, `randomForest`, and/or `ranger` packages to fit random forest models.
- `Live coding!`

Section 3

Voter File

Statara Solutions Voter File Exploration

First, who can recap what we learned about the voter file?

Statara Solutions Voter File Exploration

First, who can recap what we learned about the voter file?

Big data, hundreds of millions of entries containing best guesses for who is registered to vote in particular states, what their demographics and voting history have been, etc.

We are going to work with a smaller subsample of the voterfile (1% sample).

`Live coding!`

Section 4

Simulations

Monte Carlo Simulations: Overview

- Monte Carlo simulations are general computational technique for investigating how well something works or what might happen under given circumstances.

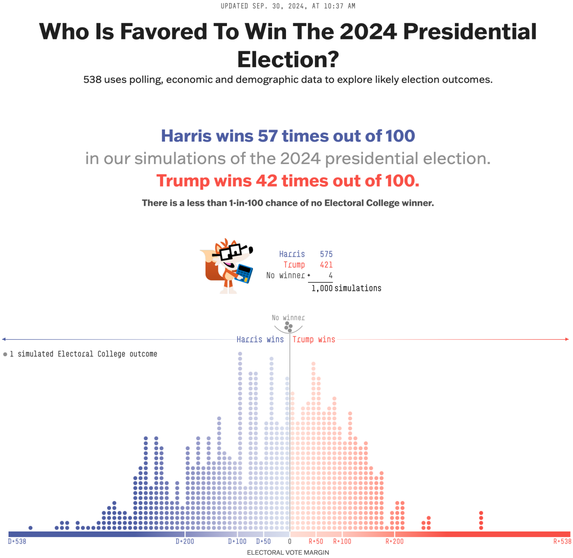
Monte Carlo Simulations: Overview

- Monte Carlo simulations are general computational technique for investigating how well something works or what might happen under given circumstances.
- Principle is repeated random sampling from given distribution (e.g., Gaussian, Uniform, etc.) based on data generating process you yourself specify.

Monte Carlo Simulations: Overview

- Monte Carlo simulations are general computational technique for investigating how well something works or what might happen under given circumstances.
- Principle is repeated random sampling from given distribution (e.g., Gaussian, Uniform, etc.) based on data generating process you yourself specify.
- An example would be simulating possible turnout and polling outcomes for 2024 for use in prediction, assessing model results accordingly.

FiveThirtyEight's Prediction



UPDATED SEP. 30, 2024, AT 10:37 AM

Who Is Favored To Win The 2024 Presidential Election?

538 uses polling, economic and demographic data to explore likely election outcomes.

Harris wins 57 times out of 100

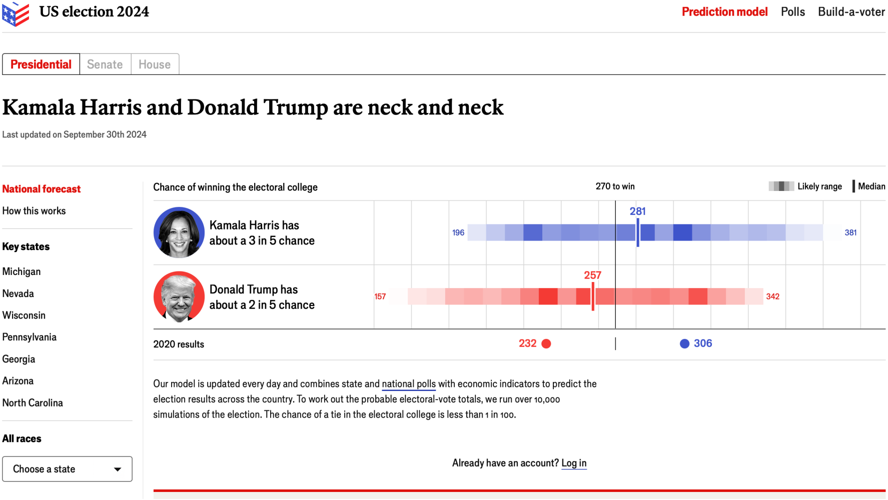
in our simulations of the 2024 presidential election.

Trump wins 42 times out of 100.

There is a less than 1-in-100 chance of no Electoral College winner.

Harris	575
Trump	421
No winner +	4
	1,000 simulations

The Economist's Prediction



US election 2024

Kamala Harris and Donald Trump are neck and neck

Last updated on September 30th 2024

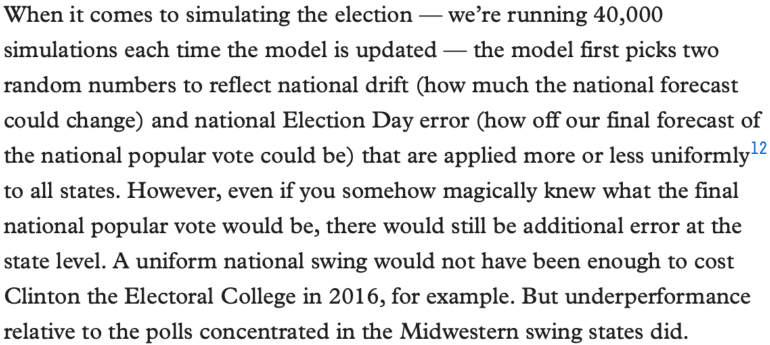
Kamala Harris has about a 3 in 5 chance

Donald Trump has about a 2 in 5 chance

Our model is updated every day and combines state and national polls with economic indicators to predict the election results across the country. To work out the probable electoral-vote totals, we run over 10,000 simulations of the election. The chance of a tie in the electoral college is less than 1 in 100.

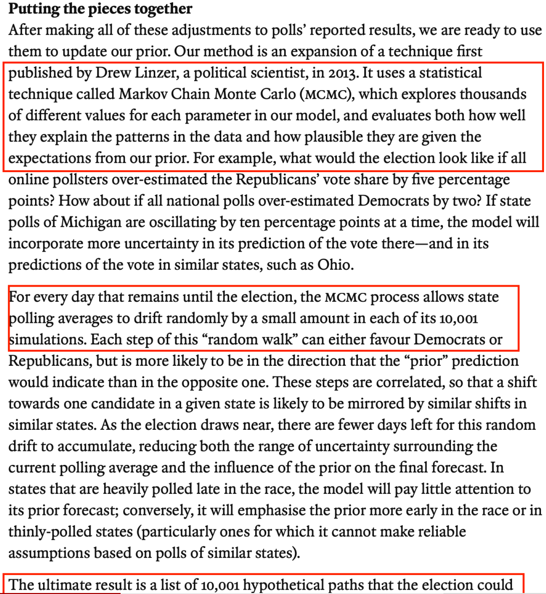
Already have an account? Log in

Silver Bulletin's Simulation Methodology



When it comes to simulating the election — we're running 40,000 simulations each time the model is updated — the model first picks two random numbers to reflect national drift (how much the national forecast could change) and national Election Day error (how off our final forecast of the national popular vote could be) that are applied more or less uniformly[12] to all states. However, even if you somehow magically knew what the final national popular vote would be, there would still be additional error at the state level. A uniform national swing would not have been enough to cost Clinton the Electoral College in 2016, for example. But underperformance relative to the polls concentrated in the Midwestern swing states did.

The Economists's Simulation Methodology



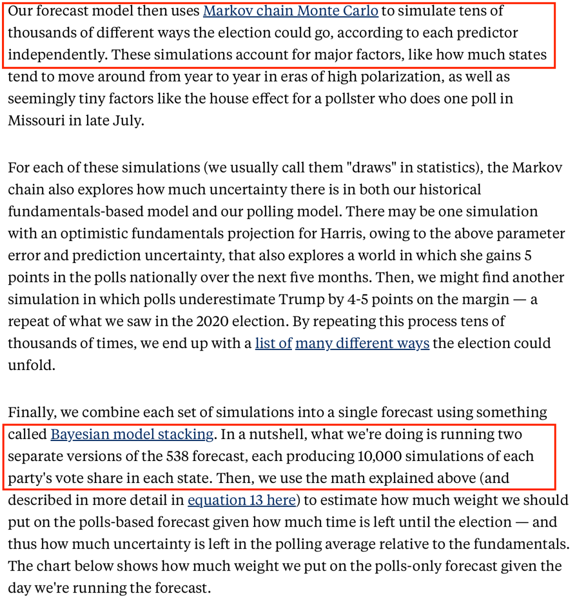
Putting the pieces together

After making all of these adjustments to polls' reported results, we are ready to use them to update our prior. Our method is an expansion of a technique first published by Drew Linzer, a political scientist, in 2013. It uses a statistical technique called Markov Chain Monte Carlo (MCMC), which explores thousands of different values for each parameter in our model, and evaluates both how well they explain the patterns in the data and how plausible they are given the expectations from our prior. For example, what would the election look like if all online pollsters over-estimated the Republicans' vote share by five percentage points? How about if all national polls over-estimated Democrats by two? If state polls of Michigan are oscillating by ten percentage points at a time, the model will incorporate more uncertainty in its prediction of the vote there—and in its predictions of the vote in similar states, such as Ohio.

For every day that remains until the election, the MCMC process allows state polling averages to drift randomly by a small amount in each of its 10,001 simulations. Each step of this "random walk" can either favour Democrats or Republicans, but is more likely to be in the direction that the "prior" prediction would indicate than in the opposite one. These steps are correlated, so that a shift towards one candidate in a given state is likely to be mirrored by similar shifts in similar states. As the election draws near, there are fewer days left for this random drift to accumulate, reducing both the range of uncertainty surrounding the current polling average and the influence of the prior on the final forecast. In states that are heavily polled late in the race, the model will pay little attention to its prior forecast; conversely, it will emphasise the prior more early in the race or in thinly-polled states (particularly ones for which it cannot make reliable assumptions based on polls of similar states).

The ultimate result is a list of 10,001 hypothetical paths that the election could

FiveThirtyEight's Simulation Methodology



Our forecast model then uses Markov chain Monte Carlo to simulate tens of thousands of different ways the election could go, according to each predictor independently. These simulations account for major factors, like how much states tend to move around from year to year in eras of high polarization, as well as seemingly tiny factors like the house effect for a pollster who does one poll in Missouri in late July.

For each of these simulations (we usually call them "draws" in statistics), the Markov chain also explores how much uncertainty there is in both our historical fundamentals-based model and our polling model. There may be one simulation with an optimistic fundamentals projection for Harris, owing to the above parameter error and prediction uncertainty, that also explores a world in which she gains 5 points in the polls nationally over the next five months. Then, we might find another simulation in which polls underestimate Trump by 4-5 points on the margin — a repeat of what we saw in the 2020 election. By repeating this process tens of thousands of times, we end up with a list of many different ways the election could unfold.

Finally, we combine each set of simulations into a single forecast using something called Bayesian model stacking. In a nutshell, what we're doing is running two separate versions of the 538 forecast, each producing 10,000 simulations of each party's vote share in each state. Then, we use the math explained above (and described in more detail in equation 13 here) to estimate how much weight we should put on the polls-based forecast given how much time is left until the election — and thus how much uncertainty is left in the polling average relative to the fundamentals. The chart below shows how much weight we put on the polls-only forecast given the day we're running the forecast.

Simulation Example

- Many different ways in which one can use simulation to help contextualize election predictions.

Simulation Example

- Many different ways in which one can use simulation to help contextualize election predictions.
- Some examples:
    - Simulate polls using random walks like in Elliot Morris' toy election simulator example.

Simulation Example

- Many different ways in which one can use simulation to help contextualize election predictions.
- Some examples:
  - Simulate polls using random walks like in Elliot Morris' toy election simulator example.
  - Simulate potential state-level turnout.

Simulation Example

- Many different ways in which one can use simulation to help contextualize election predictions.
- Some examples:
  - Simulate polls using random walks like in Elliot Morris' toy election simulator example.
  - Simulate potential state-level turnout.
  - Simulate demographic distributions.

Simulation Example

- Many different ways in which one can use simulation to help contextualize election predictions.
- Some examples:
    - Simulate polls using random walks like in Elliot Morris' toy election simulator example.
    - Simulate potential state-level turnout.
    - Simulate demographic distributions.
    - Simulate all of the above.
- `Live coding!`

Breakout Exercise

Work on building simulations for your own models and see how many times each candidate wins based on ranges of plausible values for your key variables in 2024.

Blog Extensions

N.B. Focus for this weeek on refining your models for national popular vote and Electoral College prediction. Feel free to use these optional extensions as potential starting points for doing so.

[1] Reference the *FiveThirtyEight* or *Economist* websites for for examples of simulation-based forecasting uncertainty.

Blog Extensions

N.B. Focus for this weeek on refining your models for national popular vote and Electoral College prediction. Feel free to use these optional extensions as potential starting points for doing so.

1. **Voter File.** Incorporate an analysis of the voter file samples in your models for this week, either using all of the states, a subset of the states, or a single state of interest.
2. **Random Forest Prediction.** Implement a random forest model to conduct a demographic analysis or prediction using replication data from the ANES as in Kim & Zilinsky (2023). Perhaps you can replicate or find errors in their code! Or use these models on state-level demographic data.
3. **Simulations.** Use simulations to quantify the uncertainty in your prediction.[1]

[1] Reference the *FiveThirtyEight* or *Economist* websites for for examples of simulation-based forecasting uncertainty.