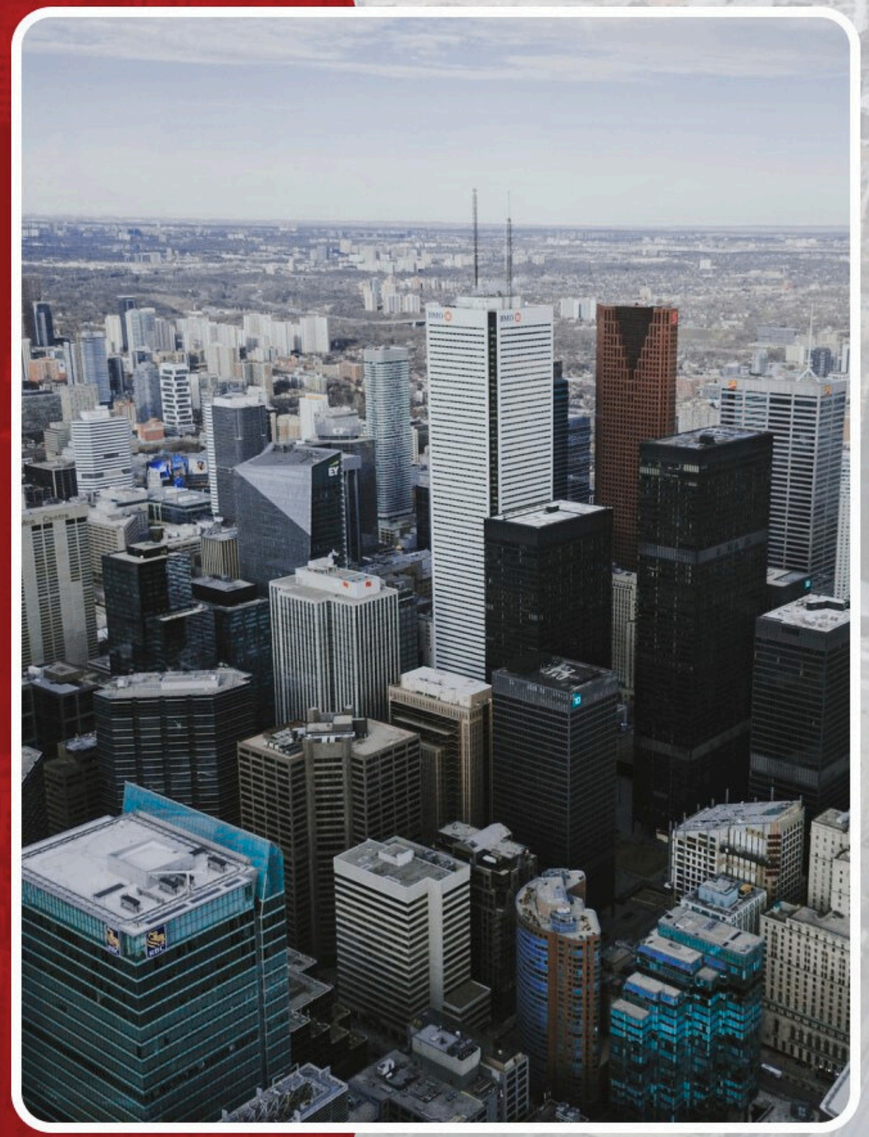
# Predicting Party Affiliation from Demographics

**Testing conventional wisdom about age, income, and gender as predictors of partisanship in New York State**

**2021 New York Voter File Analysis**

**n ≈ 84,000 registered voters | 1% representative sample**

By: Samari, Matthew, and Ainsley



# What Conventional Wisdom Says

**In political science and polling, demographic variables like age, income, and gender are widely treated as strong, reliable predictors of party identification. These factors reflect generational political socialization patterns and create predictable voting blocs observed across multiple electoral cycles. But does this hold when we test it on real individual-level data?**

## Age

## Income

## Gender

# Why New York?

## Internal diversity

While New York votes solidly Democratic in aggregate, it contains significant regional and demographic variation - making it an ideal test case where patterns aren't obvious at first glance.

## Individual-level analysis

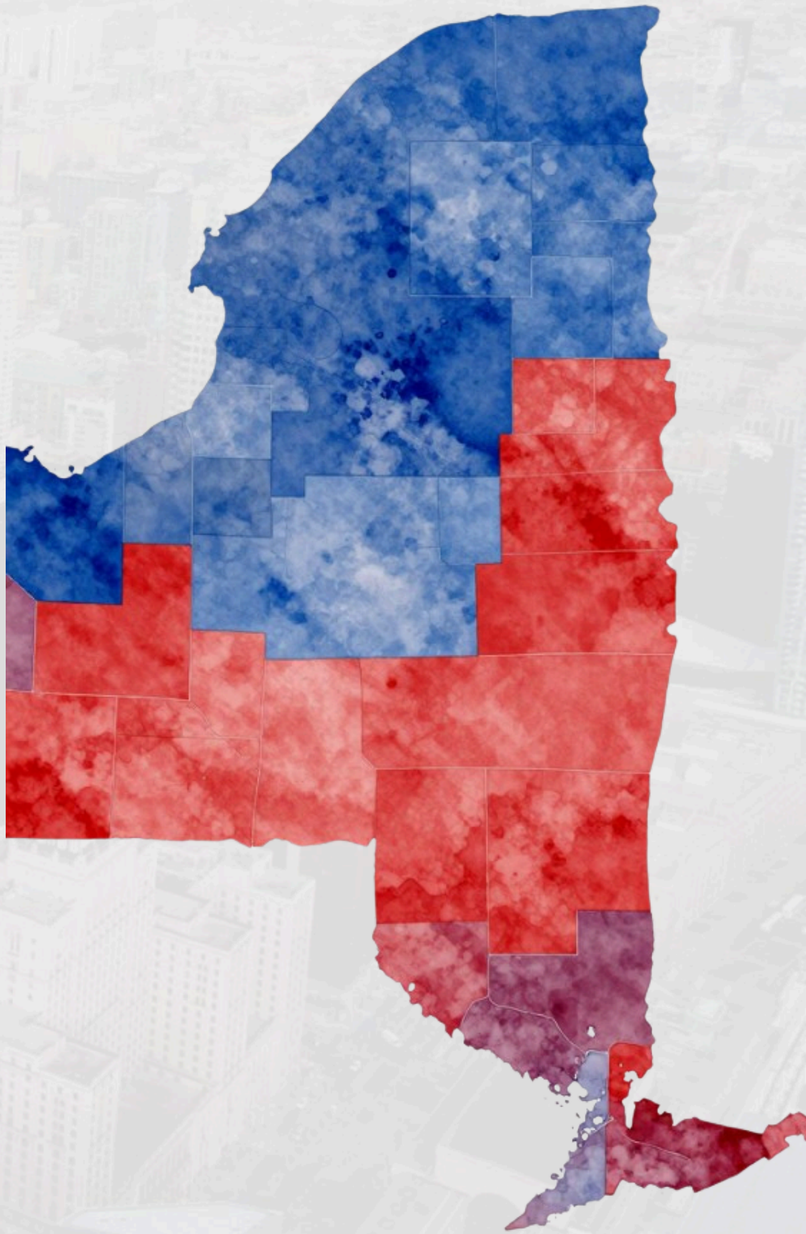
We analyze actual voter behavior at the individual level, cutting against typical state-level polling that often masks important micro-level variations.

## Real data

Our findings are based on real voter records and actual behavior, not surveys or predictive models - providing concrete evidence rather than projections.

# Data: 2021 New York Voter File

**L2 2021 New York Voter File (1% sample, N ≈ 84,000 voters)**



## 1 Exclusions

- **Inactive & non-partisan voters excluded for simplicity**
- **Education/ethnicity data excluded due to quality issues**

# Our Hypothesis

## 1. Income

Hypothesized as strongest predictor of party affiliation based on socioeconomic status theories

## 2. Age

Hypothesized as moderate predictor due to generational political socialization patterns

## 3. Gender

Hypothesized as weakest predictor given recent convergence in political attitudes

# The Analytical Journey

1 | 2 | 3 | 4

## Raw Data

84,000 voters from NY file

## Define Variables

Age, Income, Gender
Processed through Redivis
Graphics through R

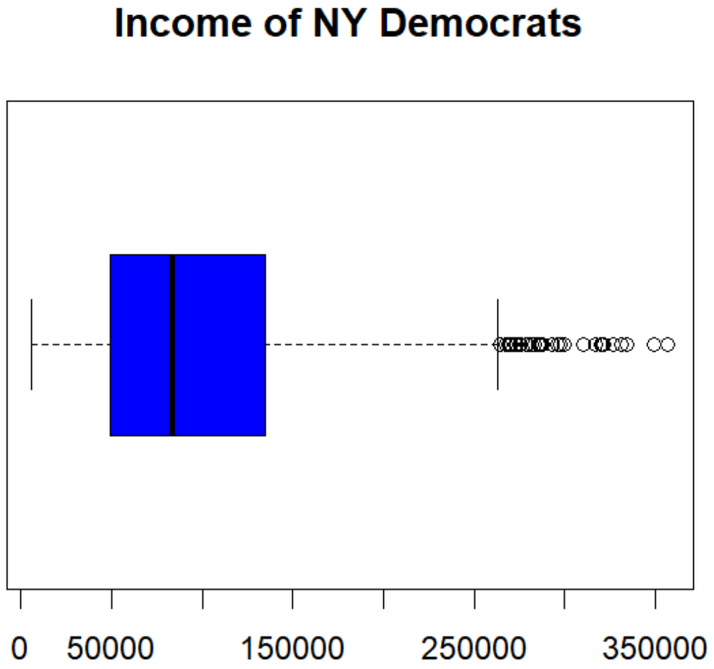
## Build Model

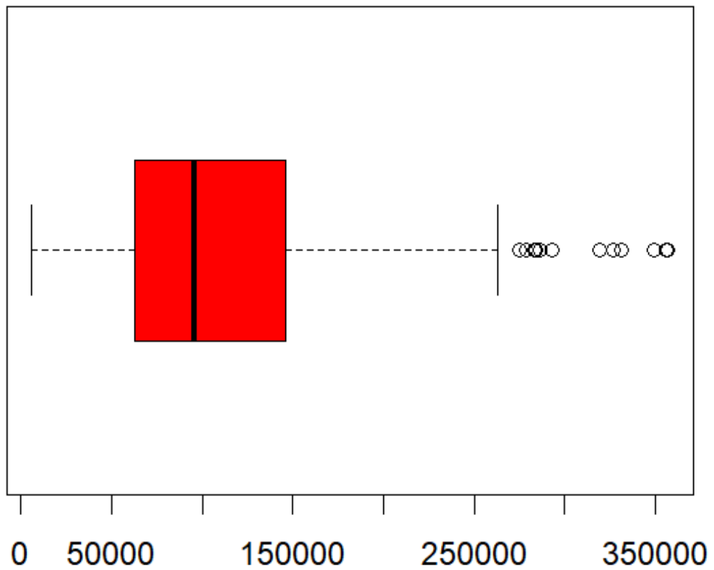
Linear regression through R

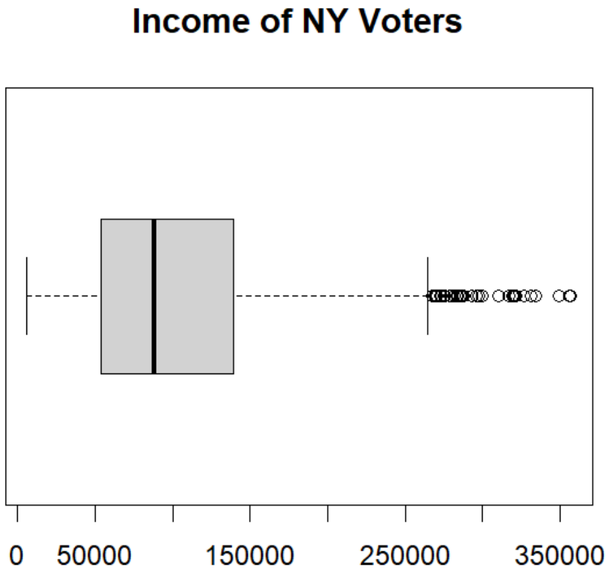
## Discover Insight

Modest predictive power

# Income Breakdown



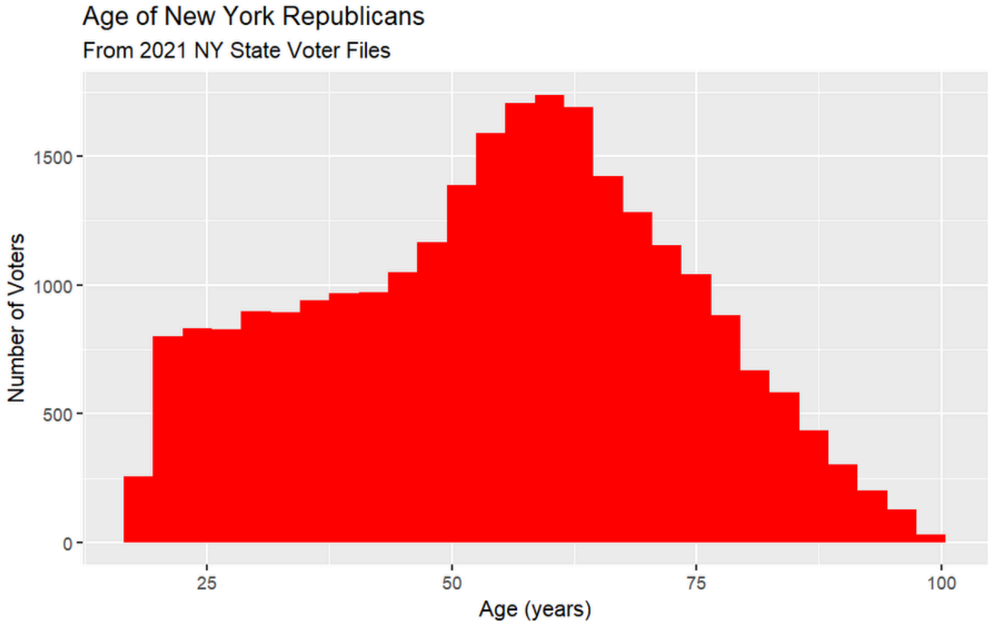

```
> rep_income <- gsub("[$,]", "", df_rep$income) |>
+   as.integer()
> summary(rep_income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6000   63000   96022  111303  146000  357020
   NA's
    430
>
> dem_income <- gsub("[$,]", "", df_dem$income) |>
+   as.integer()
> summary(dem_income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6000   49000   83954   99564  135000  357020
   NA's
   1414
>
> total_income <- gsub("[$,]", "", df$income) |>
+   as.integer()
> summary(total_income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6000   53701   88000  103182  139235  357020
   NA's
   1844
```

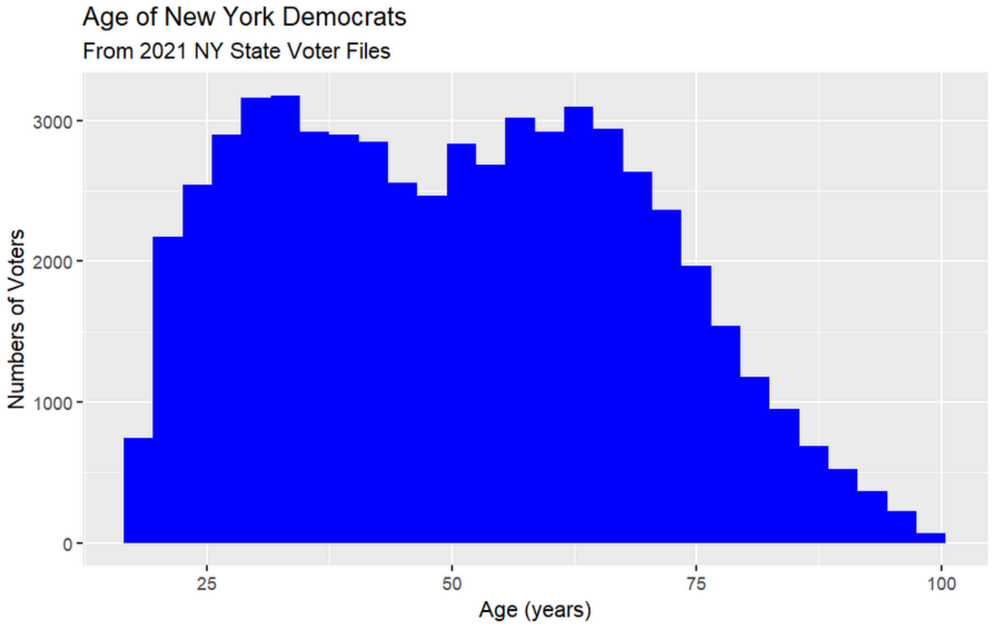
**Important to note:**

**-trends in implicit form**

**-outliers**

# Age Breakdown



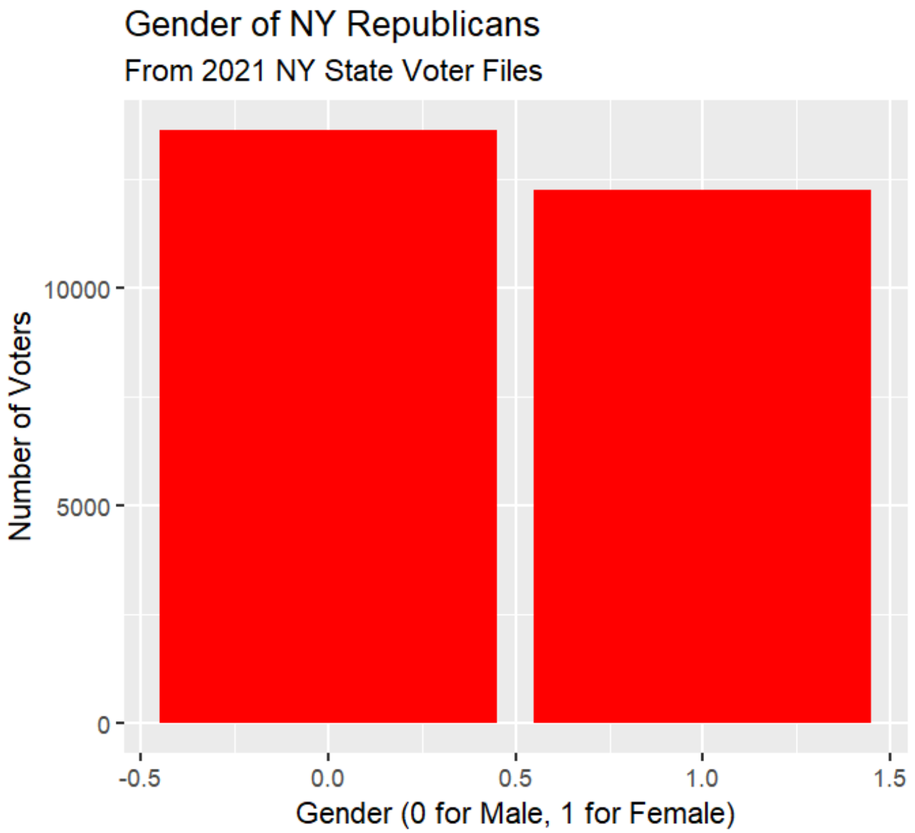



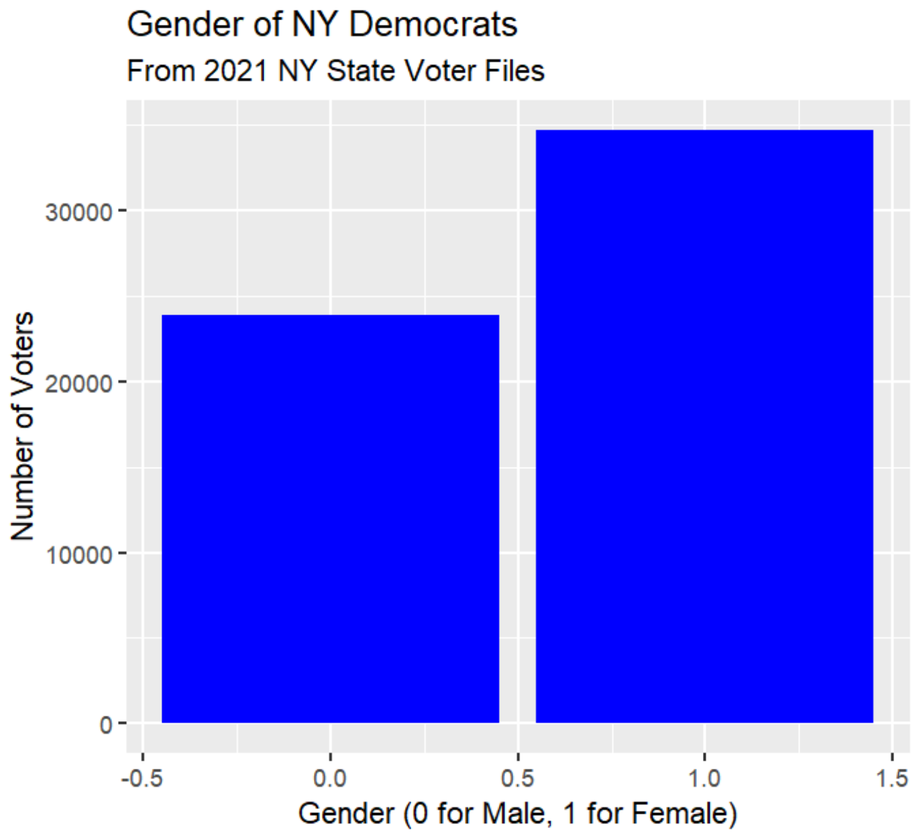

**note the scale!**

```r
# Forming a histograms for Dem and Rep with respect to age
# Important to note the scale for all of these graphs

ggplot(
  data = df_dem,
  mapping = aes(x = age)
) +
  geom_histogram(binwidth = 3, fill = "blue") +
  labs(
    title = "Age of New York Democrats",
    subtitle = "From 2021 NY State Voter Files",
    x = "Age (years)",
    y = "Numbers of Voters"
  )

ggplot(
  data = df_rep,
  mapping = aes(x = age)
) +
  geom_histogram(binwidth = 3, fill = "red") +
  labs(
    title = "Age of New York Republicans",
    subtitle = "From 2021 NY State Voter Files",
    x = "Age (years)",
    y = "Number of Voters"
  )
```




# Gender Breakdown

**Important to note:**

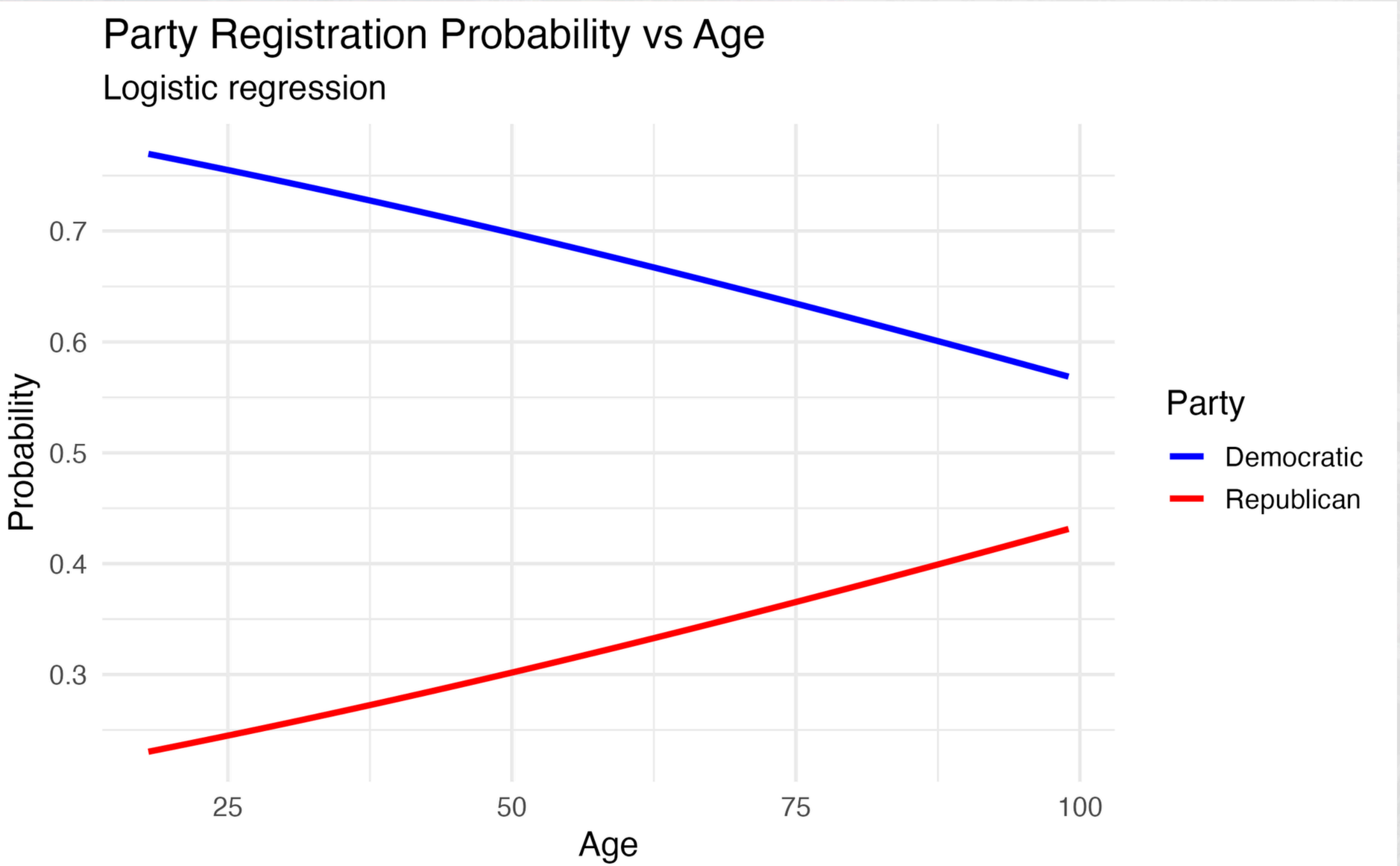
**-Binary scale for gender was done in Redivis**

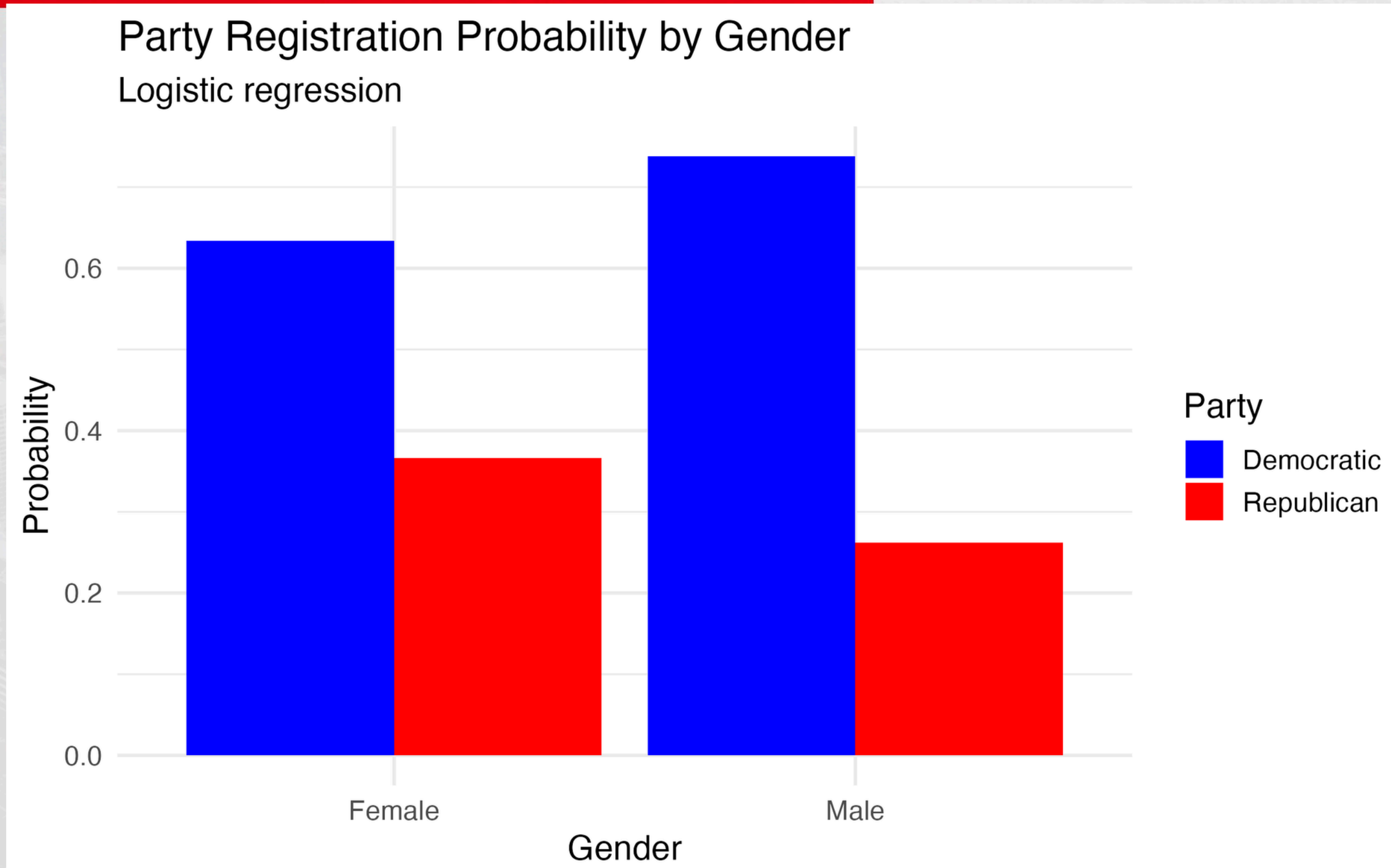
**-Gender gap**

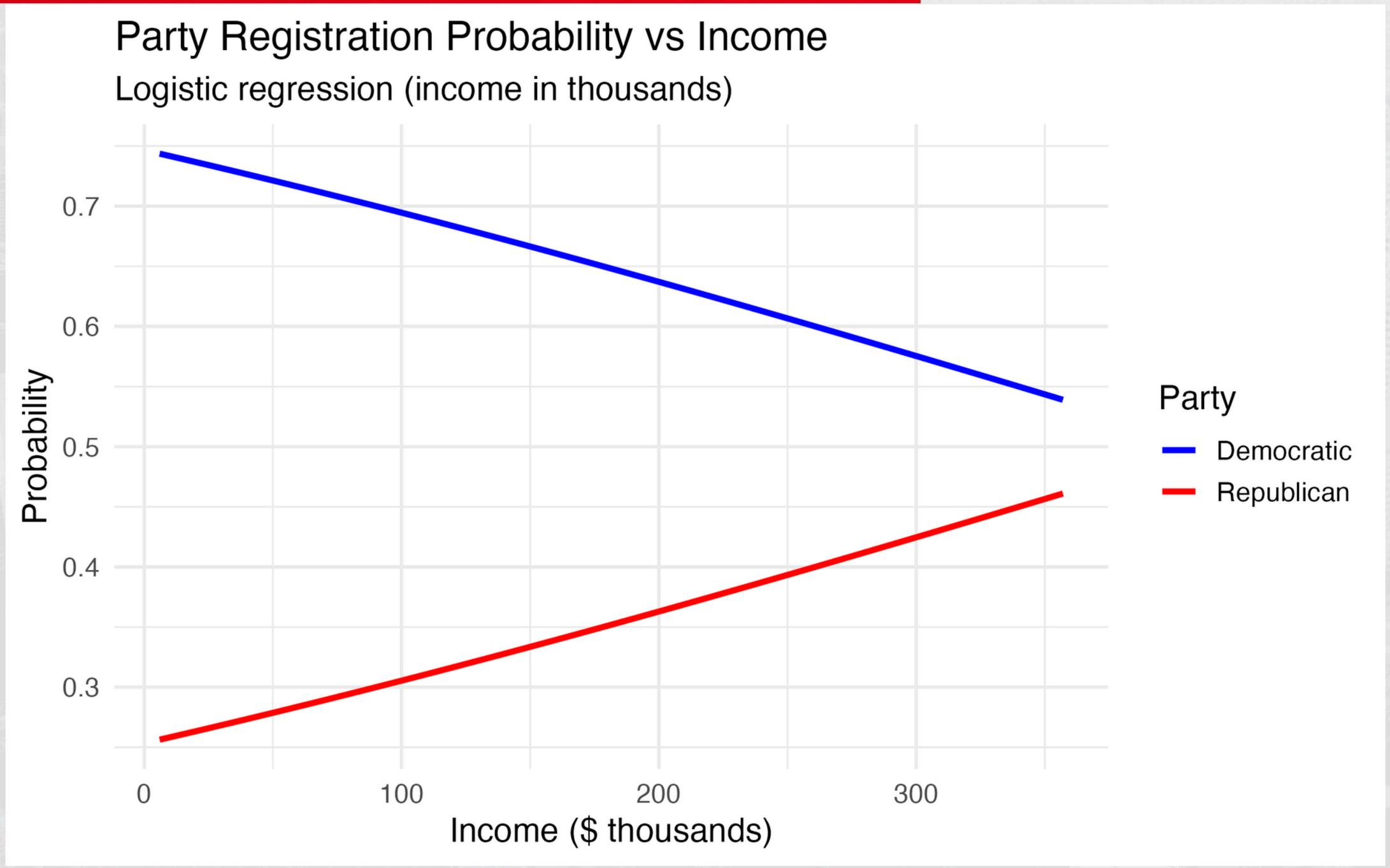
**-Scale**

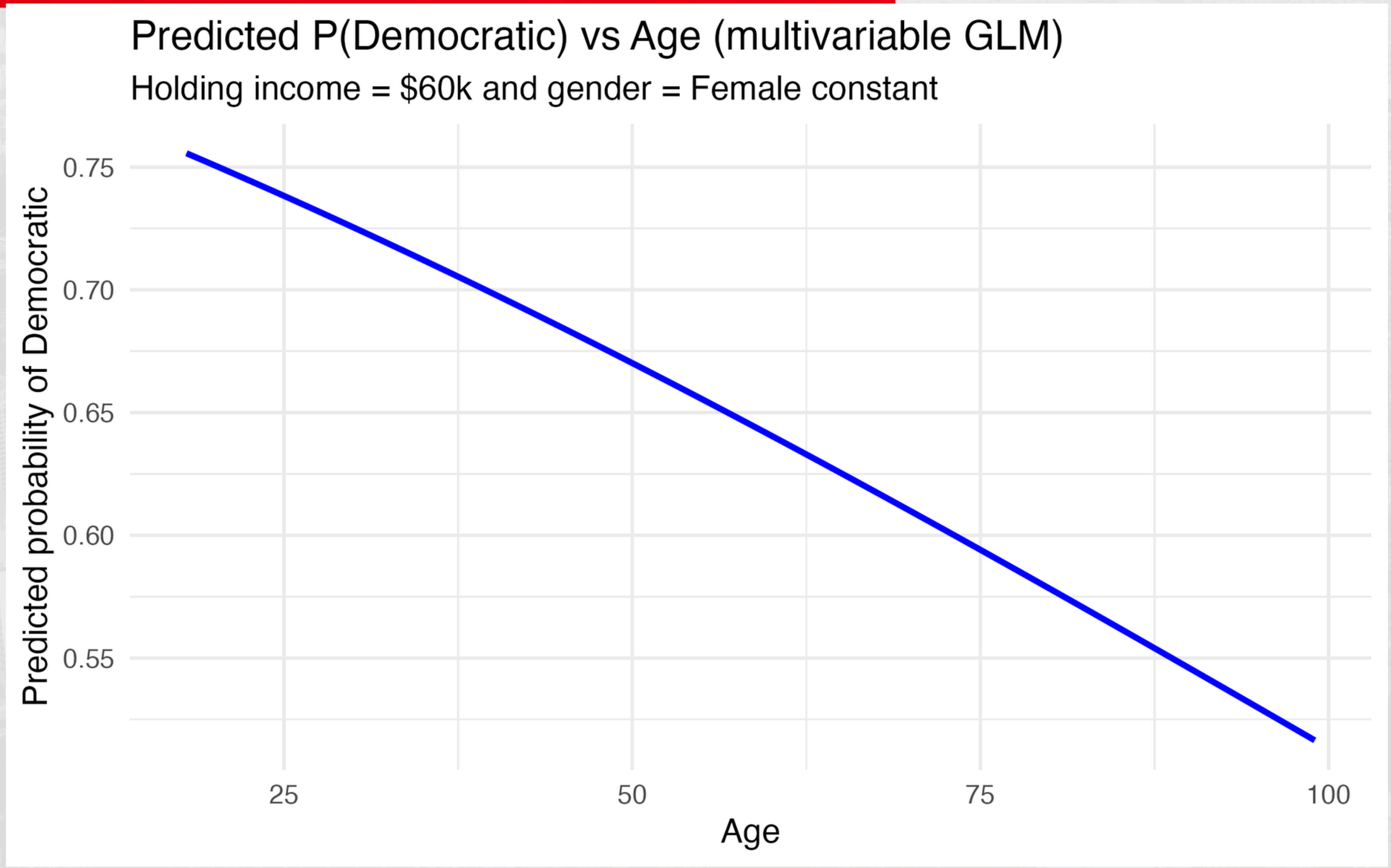
# The Model We Ran

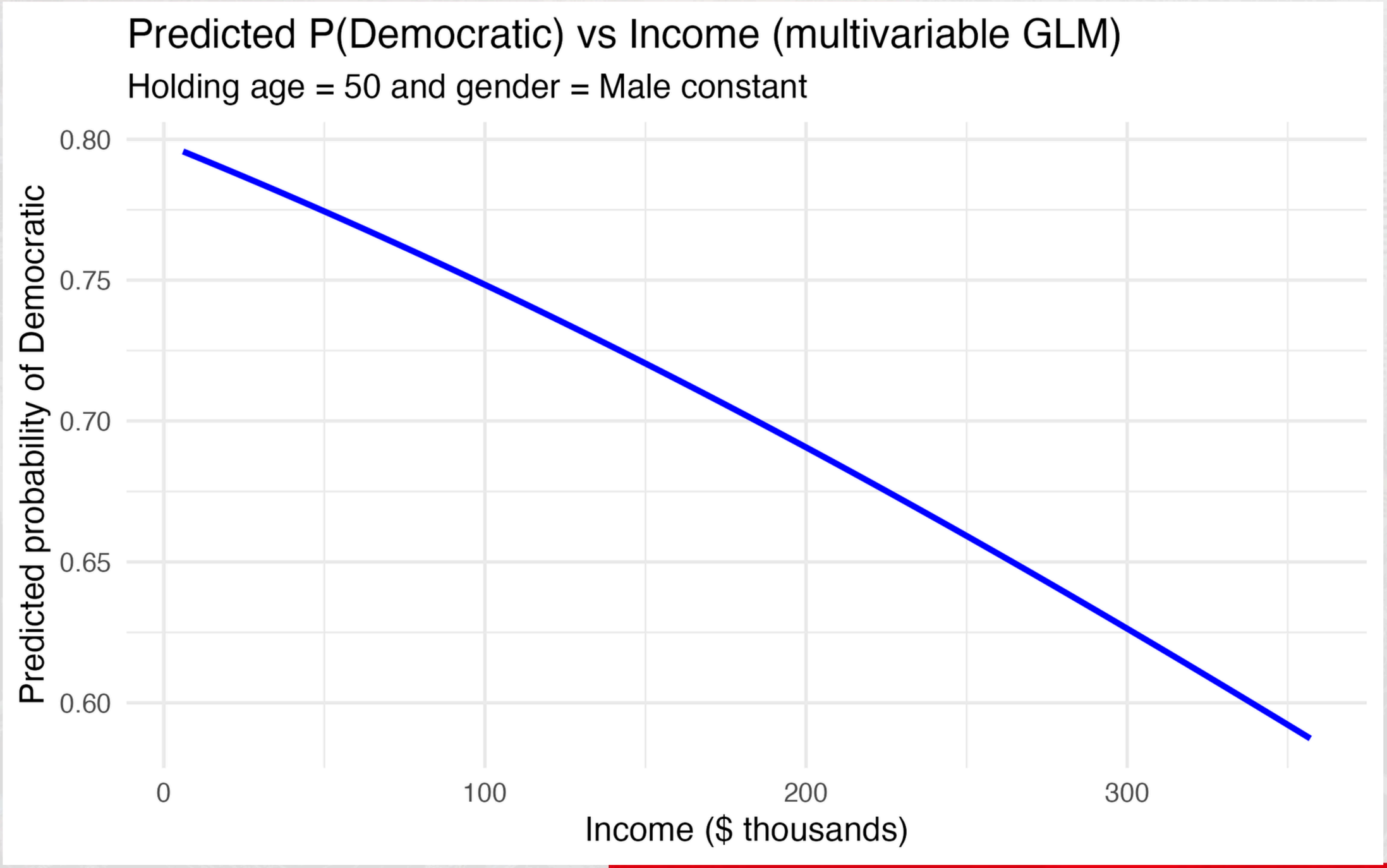
- Outcome:
  - Party registration (Democratic vs Republican)
- Method:
  - Logistic regression (Generalized Linear Model)
- Predictors:
  - Age, Income, Gender
- All predictors included simultaneously


Party Registration Probability vs Age
Logistic regression
Probability
0.7
0.6
0.5
0.4
0.3
25
50
75
100
Age
Party
Democratic
Republican


Party Registration Probability by Gender
Logistic regression
Probability
0.6
0.4
0.2
0.0
Female
Male
Gender
Party
Democratic
Republican


Party Registration Probability vs Income
Logistic regression (income in thousands)
Probability
0.7
0.6
0.5
0.4
0.3
0
100
200
300
Income ($ thousands)
Party
Democratic
Republican


Predicted P(Democratic) vs Age (multivariable GLM)
Holding income = $60k and gender = Female constant
Predicted probability of Democratic
0.75
0.70
0.65
0.60
0.55
25
50
75
100
Age


Predicted P(Democratic) vs Income (multivariable GLM)
Holding age = 50 and gender = Male constant
Predicted probability of Democratic
0.80
0.75
0.70
0.65
0.60
0
100
200
300
Income ($ thousands)

# Why This Matters: Rethinking Voter Prediction

## Generalization is risky

State-level or national averages obscure individual variation

## Hidden factors

Education, religion, family history, cultural identity likely matter more than simple demographics

## Complexity matters

Treating all Democrats or Republicans as a homogeneous demographic bloc is fundamentally flawed

# Limitations and Future Work

## Limitations

Excluded 26K non-partisan voters

Limited to 2021 snapshot

Missing key variables (education, religion, race/ethnicity)

## Future Directions

Multi-state comparison

Incorporate unmeasured variables via survey

Explore non-partisan voters separately