Fact versus Fiction: Verifying Predictor Conventions for New York

Matthew Cohen[*] Ainsley Hoover[†] Samari Ijezie[‡]

2026-01-10

Abstract

This study examines how well age, income, and gender predict party identification among registered voters in New York State. Using a 1% sample of the 2021 New York voter file from L2 (N 84,000), we estimate the predictive power of these demographic variables using bivariate and generalized linear models. We find that income and age provide modest predictive value for distinguishing between Democratic and Republican affiliation, while gender contributes comparatively little explanatory power. We made a predictive model that predicts probabilities of party registration given age, gender, and income. We caution against using this data for individuals, instead focusing on communities at large. These findings suggest that while demographic trends exist in aggregate, party identification in New York has apparent correlations with income, age, and gender.

1 Introduction

Voting plays a central role in democratic societies by allowing individuals to influence who governs and which policies are prioritized. In the United States, presidential elections are especially consequential, shaping national leadership, economic policy, social programs, and judicial appointments. One of the earliest and most formal expressions of political participation is party registration, which reflects how individuals align themselves within the political system and determines eligibility to participate in party primary elections.

New York State provides a useful setting for examining party registration. As one of the largest states in the country, New York is often described as reliably Democratic, yet it includes substantial political, geographic, and socioeconomic diversity. Party affiliation also carries particular importance in New York because the state operates under a closed primary system. Voters must be registered with a political party in advance in order to vote in that party's primary elections. While individuals are allowed to change their party registration, they must do so before established deadlines. As a result, party registration in New York represents a deliberate and meaningful signal of political identification rather than a temporary or symbolic choice.

Political science research and public discussion frequently assume that demographic characteristics such as age, income, and gender are strong predictors of political identification. These assumptions

[*] American University
[†] American University
[‡] American University

shape polling, campaign strategy, and media narratives, but they are often based on broad population trends rather than tested using individual level data. This raises the question of whether these widely accepted beliefs about voter behavior hold when examined more closely.

Throughout our paper, we use active, first-person language and avoid the passive voice.Specifically, we estimate the probability that an individual registers as Democratic or Republican given their age, gender, and income. Using individual level voter file data and logistic regression models, we evaluate the extent to which basic demographic factors explain variation in party registration. For example, we write “we examine the relationship between *X* and *Y*”; we do not write “the relationship between *X* and *Y* was examined.” Where we do the analysis, we speak about it transparently. We use the present tense; for example, “In this paper, we argue …” and “Paper XYZ demonstrates the relationship between …”.

2 In Search of Nuance

Age, gender, and income, are commonly described as predictors for party affiliation. In introductory-level civics courses, students are taught about how the older and richer one is, the more conservative they are likely to be. As people get older, they shift away from the Republican party (Knoke and Hout 1974). In addition, the gender gap in the United States’ two party system is highly referenced. Women lean more democratic than men (Box-Steffensmeier, De Boef, and Lin 2004). This has been the rhetoric for decades (Miller 1991). There is a plethora of research showcasing the effect of demographic information on voting (a quick search on Google Scholar of “demographic effects on voting” renders more than 1.3 million results). However, attention to voting predictors is rarely given at the state level. Our country that is hyper-fixated on presidential elections. Thus, we hoped to move away from that narrative and dissect one aspect of voters at the state-level. We wanted to verify if age, gender, and income are strong determiners for categorizing partisanship or if the raw data presents a different picture.

Selecting New York as a key state may seem odd, as the state is not a swing state and heavily favors the Democratic party. In other words, nationally, it is often predictable and not a good representation of the United States of America at large. This is precisely why we chose New York: it might be an outlier to general conventions. There may be general cohort effects to the elderly or youth specific to New York that national data misses. New York City’s 2025 mayoral election elected Mamdani, an untraditional Democrat, to their office. While New York City is by no means representative of the state as a whole, it goes to show that the region acts as an anomaly to what the nation would typically elect. Thus, we wondered if New York had unique attributes, related to age, income, or gender, that allowed for its democratic-party favoritism. Age may not be the only factor at play for party identification: income and gender may hold key influence as well.

In a diverse and expansive world, it seems pertinent to take caution to generalizations. Generalizations about party identification are most accurate at the population level. It is vital to remember that the population is distinct from an individual. There is much intellectual debate about the reliability of party identification assumptions. There is some evidence that demographics can only prove partisanship by so much (Tomkins et al. 2025). In addition, there is doubt placed on models predicted partisanship. By one estimate, machine learning models trained on demographic labels from public opinion surveys predicted partisan identification correctly only 63.4% of the time (Seo-young Kim and Zilinsky 2024). By focusing on one state, we hope to shift away from nationally-applied partisanship influences and verify their accuracy for the state of New York.

3 Data and Methods

For this report, we analyzed data from the 2021 New York voter files. In it were registered voters' names, registered addresses, ages, vote histories, and party affiliations, among hundreds of other variables. We obtained this data from L2, a gold-standard database for the United States' voter files. We received this data through Joshua Ferrer's access. He gave us a 1% of the raw data, which is still substantially large enough to do successful analysis with. Our sliced data still contained 125,046 voters with 1,180 variables about them. This data is very rich in content, but some variables lacked enough presence to be useful. For example, the education data is an estimate from consumer-sold data. Despite this information being wildly circulated online, half of our data's New Yorkers were lacking education attainment level data. For this reason, it did not feel as a representative metric to use. This led us to analyzing age, gender, and income as over 90% of voters had that information.

Since this data was very large, we took many steps to make it useful for our purposes in predicting partisanship. First, we chose our useful variables of voter identification, age, date of birth, partisan affiliation, ethnicity, religion, estimated income, education, gender, activity of the voter, and the county in which the voter resided. Due to concerns about the quality of data, as many of these variables are faulty at the individual level, we removed education, religion, and ethnicity. For our goal of predicting party identification, we also filtered out inactive voters as they are not likely to have strong identification with their party affiliation and no longer vote in the state of New York. This could have been due to residency, fatality, criminal status, or otherwise. We also coded gender with a binary scale, with 0 as male and 1 as female. There were 36 voters with no gender information, which we elected to filter out, as they were not a significant portion of the population. For party identification, we decided to focus only on the affiliated Republicans and Democrats for simplicity in our data. We made all of these changes through a workspace in Redivis as to not overwhelm R.

Our primary research interest is the effectiveness of three variables, gender, age, and income, in predicting party identification. We did this through linear regression and multinomial predictors. We did the linear regression using R, relying on the tidyverse package and ggthemes. We hypothesized that income would be the strongest indicator, followed by age, and then gender.

4 Conventions Hold True

Here, we explain and interpret our results. We try to learn as much as we can about our question as possible, given the data and analysis. We present our results clearly. We interpret them for the reader with precision and circumspection. We avoid making claims that are not substantiated by our data. We are careful about causality. When we describe associations, we avoid language like "effects" and "increases"; we only describe "effects" or "impacts" when we have a causally well-identified research design.

Note that this section may be integrated into Section 3, if joining the two improves the overall presentation.

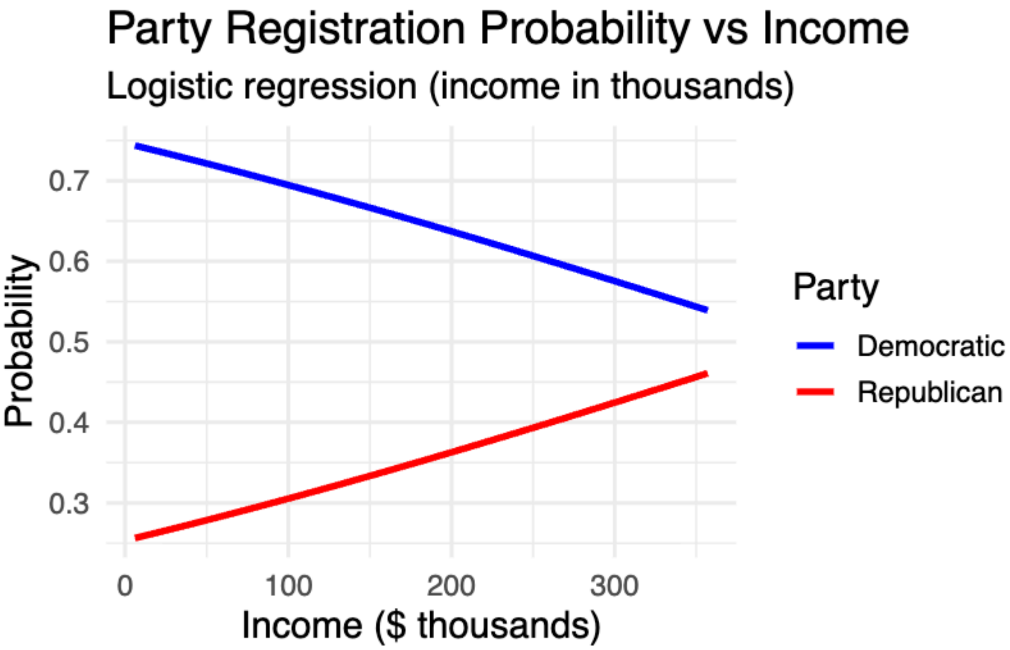
5 Results

```
Democratic Republican
     56983      25402
```

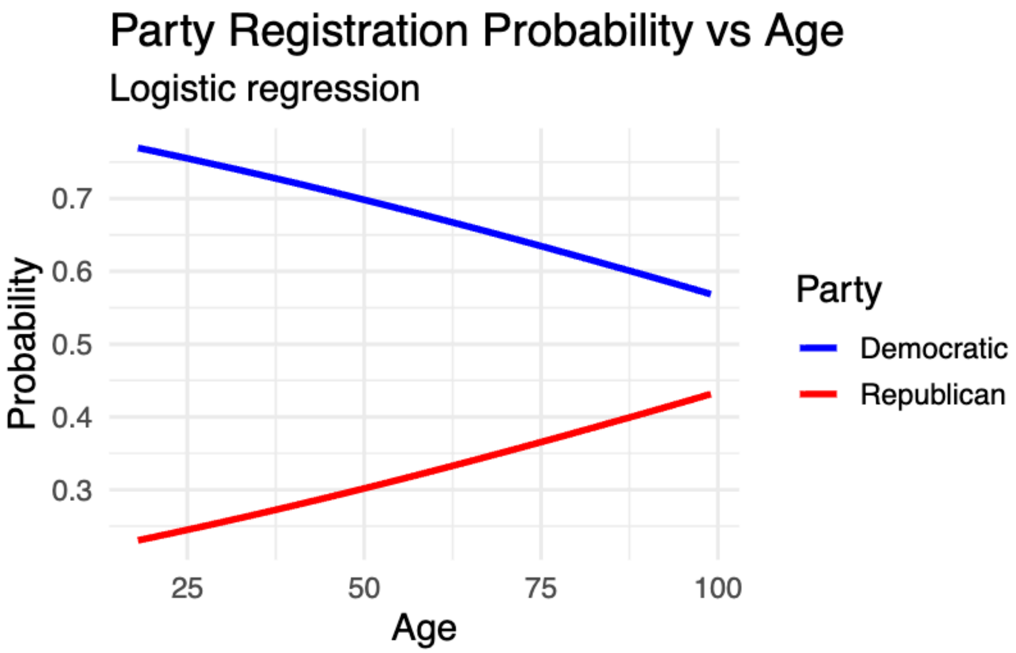
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   36.00   52.00   51.93   66.00   99.00

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   53.78   88.00  103.20  139.33  357.02
```

```
Female   Male
 36554  45831
```

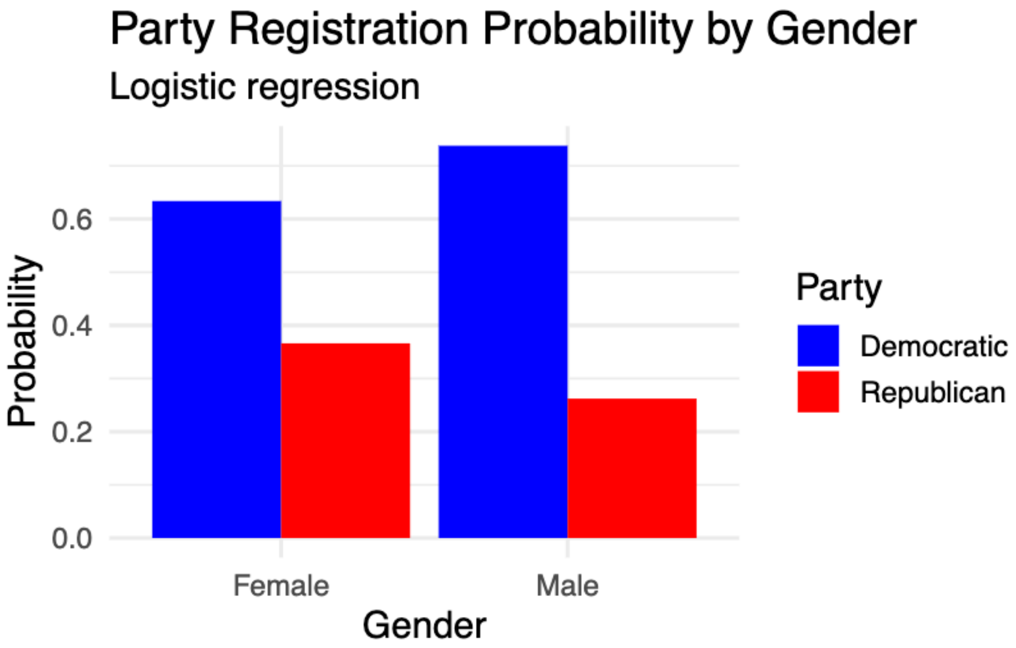
The sample includes substantially more registered Democrats (56,983) than Republicans (25,402), reflecting New York's broader partisan landscape. Voters span a wide age range from 18 to 99, with a median age of 52, while estimated income shows considerable variation and a right-skewed distribution driven by a smaller number of high-income individuals. Together, these patterns establish important baseline differences in party composition and demographic spread that inform how age and income are interpreted in the modeling results.



This chart shows how predicted party registration varies with income based on a logistic regression model. As income increases, the probability of registering as Democratic steadily declines, while the probability of registering as Republican increases. Although higher-income voters in New York become more likely to register Republican as income rises, Democrats remain more likely than Republicans across the entire income range shown, indicating that income influences party registration but does not fully overturn the state's overall Democratic lean.

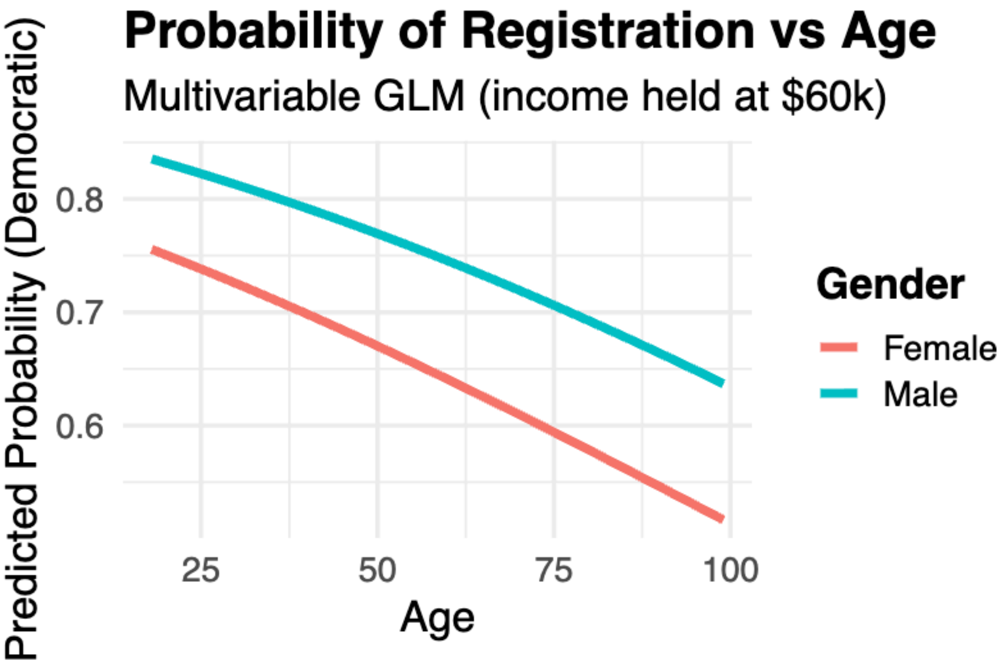


This chart shows a clear generational pattern in party registration. Younger voters are far more likely to be registered Democrats, while Republican registration steadily increases with age. Even so, Democratic affiliation remains dominant across all age groups, highlighting how age shifts partisan balance without fully reversing it.



The chart shows that both men and women in New York are more likely to be registered as Democrats than Republicans. Men have a higher predicted probability of Democratic registration than women, while women have a higher probability of Republican registration compared to men.

Overall, gender is associated with differences in party registration, but the gaps are modest rather than decisive.



This chart shows that as voters age, they become less likely to be registered as Democrats, even when income is held constant. The decline appears for both men and women, indicating that age is a strong predictor of Democratic registration across genders. While men consistently have a slightly higher probability of Democratic registration than women at each age, the downward trend with age is similar for both groups.

Logistic Regression Results

All predictors are statistically significant at the 0.1% level.

Model Fit Statistics

Null Deviance: 101,788 (df = 82,384)

Residual Deviance: 99,252 (df = 82,381)

AIC: 99,260

---

6 Discussion

Our paper adds to the large amount of literature dedicated to predicting party affiliation. We find that common rhetoric related to how age, income, and gender impact partisanship holds true for New York. As New York voters age, they become more likely to be Republicans. The same is true for if they are wealthier. Gender had unique attributes that varies on a voter's estimated income and age. This work is important as it demonstrates that despite New York's abundance of

Table 1: Our Informative Title

	Outcome
	dem
age	-0.01^{***}
	(0.0004)
income_k	-0.003^{***}
	(0.0001)
genderMale	0.50^{***}
	(0.02)
Constant	1.54^{***}
	(0.03)
Observations	82,385
Log Likelihood	−49,626.16
Akaike Inf. Crit.	99,260.32
Note:	*p<0.1; **p<0.05; ***p<0.01

Democrats, the traditional demographic trends hold true. It demonstrates that, at least for New York, national-scale research is sound at the state-level.

One limitation was our inability to involve third-party and non-partisan affiliations into our analysis. We recommend future work to be done that incorporates non-partisan affiliates, as they represented a large portion of the New York Voter File (26,108 voters of our 116,801 data set were non-partisan). In addition, other factors than the three we tested may be more responsible for party affiliation, such as education, party identification of parents, and race/ethnicity. The voter files alone did not contain a high quality measure, or one at all, for these potential predictors. We suggest the use of a survey to view correlation between these predicted determiners. A natural expansion of this project would be to do the same methodology for many states. Then, comparison between regions or different states could render intriguing findings. We were also limited by our access to the voter files. Were a more recent file be available, such as 2025, that would better encapsulate the voting dynamics of today's world than 2021. That being said, we expect limited changes for party identification as affiliation is rarely updated in official records. Additionally, it is likely that other predictors within our data set will glean interesting information related to partisanship. We chose predictors that are commonly described as strong influences, but it is possible some of the hundreds of other variables demonstrate correlations.

7 Conclusion

To conclude, the data largely supports public perception regarding partisan identification in New York State. Age and income emerge as consistent predictors of party registration, with individuals becoming less likely to be registered as Democrats as either variable increases. These patterns appear across both bivariate and multivariable models, reinforcing the idea that commonly cited demographic trends do reflect meaningful differences in political alignment at the individual level.

The most notable departure from public belief appears in the role of gender. While women are often assumed to lean more Democratic, the results suggest that women in this sample are slightly more likely to be registered as Republicans once age and income are taken into account. Although the magnitude of this difference is modest, it challenges simplified narratives about gender and partisanship and suggests that gender effects may be more context-specific than widely assumed.

Taken together, these findings show that demographic characteristics shape the direction of party registration but do not fully determine it. Age and income provide strong signals about partisan tendencies, while gender plays a more limited and nuanced role. This highlights both the usefulness and the limits of demographic explanations and underscores the importance of moving beyond surface-level assumptions when interpreting patterns in voter behavior.

References

Box-Steffensmeier, Janet M., Suzanna De Boef, and Tse-Min Lin. 2004. "The Dynamics of the Partisan Gender Gap." *American Political Science Review* 98 (3): 515–28. https://doi.org/10.1017/S0003055404001315.

Knoke, David, and Michael Hout. 1974. "Social and Demographic Factors in American Political Party Affiliations, 1952-72." *American Sociological Review* 39 (5): 700–713. https://doi.org/10.2307/2094315.

Miller, Warren E. 1991. "Party Identification, Realignment, and Party Voting: Back to the Basics." *The American Political Science Review* 85 (2): 557–68. https://doi.org/10.2307/1963175.

Seo-young Kim, Silvia, and Jan Zilinsky. 2024. "Division Does Not Imply Predictability: Demographics Continue to Reveal Little About Voting and Partisanship." *Political Behavior* 46 (1): 67–87. https://doi.org/10.1007/s11109-022-09816-z.

Tomkins, Sabina, David Rothschild, Alex Liu, and Alexander Thompson. 2025. "Identity Isn't Everything – How Far Do Demographics Take Us Towards Self-Identified Party ID?" arXiv. https://doi.org/10.48550/arXiv.2507.06193.