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. Overall, demographic variables alone perform weakly at the individual level, highlighting the limits of commonly cited partisan generalizations. 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

In this section, we introduce the reader to the phenomenon we investigate. We describe the way in which our analysis contributes to an important intellectual debate, or how it answers a pressing political or social question. We introduce our research question, hypotheses, data, and results. We signpost for the reader what's coming in the rest of the paper.

We remember that our paper is not a mystery novel. We note our core results early and often.

Throughout our paper, we use active, first-person language and avoid the passive voice. 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

*American University
†American University
‡American University

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 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 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.

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.