Demographic Determinants of U.S. Presidential Election Outcomes: A County-Level Analysis of the 2020 Vote

Ahmad Khan, Carter Pettid

2025-05-07

1. Introduction

Understanding the intricate relationship between demographic characteristics and voting patterns is fundamental to comprehending political behavior and forecasting election outcomes. Factors such as age, race, income, and education level have long been focal points of political science research, with numerous studies suggesting their significant, albeit evolving, roles in shaping voter preferences and electoral results. For instance, research from the Pew Research Center on the 2022 elections highlighted distinct voting preferences among older voters and higher-income groups. However, the precise predictive power of demographics remains a subject of debate. Some scholarship, such as studies from American University, posits that demographic variables may not be as determinative of election outcomes as commonly assumed, thereby adding layers of complexity to such analyses.

This project aims to contribute to this ongoing discussion by conducting a detailed analysis of comprehensive datasets. We integrate the 2020 U.S. Census data with historical election data spanning from 2010 to 2020, and prospectively including 2024 data as it becomes available. The primary objective is to model the relationships between key demographic variables and election results at the county level. By doing so, this research seeks to enhance our understanding of how these factors influence electoral outcomes and to improve the predictive modeling of elections, especially in light of evolving demographic landscapes shaped by changes in income distribution, unemployment rates, educational attainment, and migration patterns.

2. Data and Methods

2.1 Data Collection

This project draws upon two principal categories of data:

- **Census Data:** Demographic information is primarily sourced from the 2020 U.S. Census, obtained via the U.S. Census Bureau's official website. For intercensal years, annual estimates are derived from the American Community Survey (ACS), ensuring that the demographic data remains current and reflective of ongoing societal changes.

- **Election Data:** Historical election data, covering the period from 2010 to 2020, has been compiled from reputable databases such as the MIT Election Lab and the U.S. House of Representatives' History, Art & Archives office. Data for the 2024 election cycle will be sourced from official state election commission websites or through aggregated data from established media outlets as it becomes available. The analysis presented focuses on the 2020 election data.

2.2 Data Merging and Cleaning

To effectively link demographic characteristics with voting outcomes, data merging was performed at the county/state level using Federal Information Processing Standard (FIPS) codes. These standard geographic identifiers are common to both census and election datasets, facilitating a robust alignment. This level of granularity was chosen as it offers a balance between detailed analysis and data availability.

During the data cleaning phase, counties with substantially incomplete records were excluded from the analysis. Sensitivity analyses were conducted to ascertain that these exclusions introduced minimal bias to the overall dataset and subsequent findings.

2.3 Handling Outliers and Data Quality

Initial data screening identified a small number of counties exhibiting outlier values for variables like median income or population size (e.g., exceptionally wealthy urban centers or sparsely populated rural regions). Each potential outlier was subjected to a rigorous validation process:

- **Contextual Check:** We investigated whether these outlier counties represented genuinely distinct demographic or economic conditions or if the data might be erroneous. For instance, extremely high or low population densities were often found to reflect legitimate megacities or remote frontier counties, rather than data errors. This was achieved by cross-referencing data with census tract maps and local economic reports. For example, a county with a median household income significantly above the national average was confirmed to be a major technology hub, while counties with very small populations were verified as genuine frontier regions.
- **Decision on Inclusion:** Outliers deemed valid (e.g., a high-income county with consistent data across multiple variables) were retained to ensure the dataset accurately reflected real-world heterogeneity. Such cases, like an affluent suburban county with a high proportion of votes for one party, provided valuable insights into nuanced voting patterns. Conversely, counties with clearly flawed or extensively incomplete data (e.g., over 30% of survey fields blank or filled with extreme IPUMS defaults) were omitted.

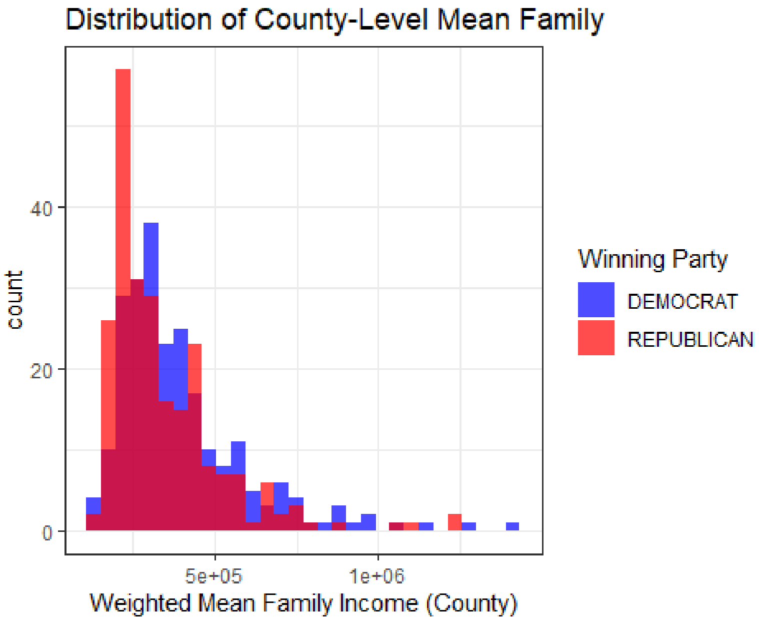
- **Dataset Fillers:** Attention was paid to how IPUMS handles missing values (NA), which sometimes involves using extreme values (like state maxima) that could distort analyses if unaddressed. To mitigate such “max-fill” distortions, filled values were compared against American Community Survey (ACS) releases. County statistics deviating by more than 20 percent from ACS figures were flagged for manual review rather than automatic inclusion.

3. Exploratory Data Analysis (EDA) and Visualization

Before developing formal statistical models, an exploratory data analysis (EDA) was conducted to summarize the data, understand variable distributions, identify notable trends, and assess potential issues such as outliers or multicollinearity. Given the complexity of election data and the often narrow margins of victory in recent elections, this initial exploration is crucial for laying the groundwork for robust modeling. The following analysis focuses on key demographic and socioeconomic trends in the 2020 election data, examining variations by county and the winning political party.

3.1 Distribution of County-Level Mean Family Income by Winning Party

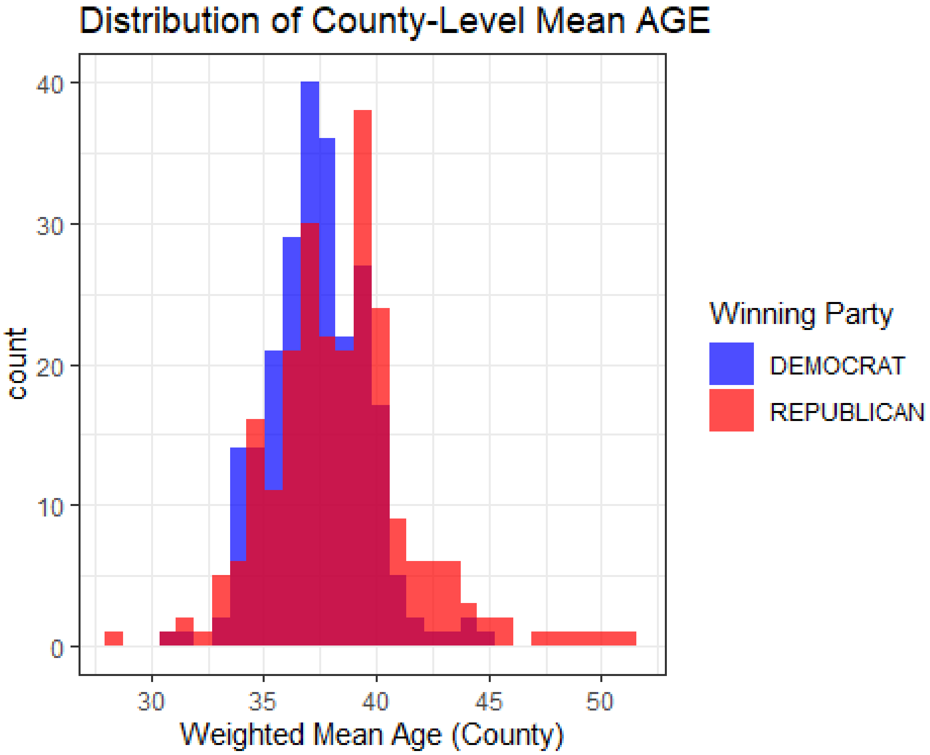
The histogram illustrating county-level weighted mean family incomes, disaggregated by the winning party (Democrat in blue, Republican in red), reveals that most counties are concentrated at lower income levels. Both parties show significant representation in counties with mean family incomes below $500,000. Notably, counties with higher mean family incomes (exceeding $500,000) appear to favor Democratic candidates more frequently, although such high-income counties are relatively uncommon.



- *Insights:* This distribution suggests a potential influence of income on voting patterns. While lower-income counties are more evenly split, higher-income counties show a tendency to lean Democratic. This observation warrants further investigation in subsequent modeling. However, the skewed nature of the income distribution and the presence of high-end outliers necessitate careful statistical treatment in further analyses.

3.2 Distribution of County-Level Mean Age by Winning Party

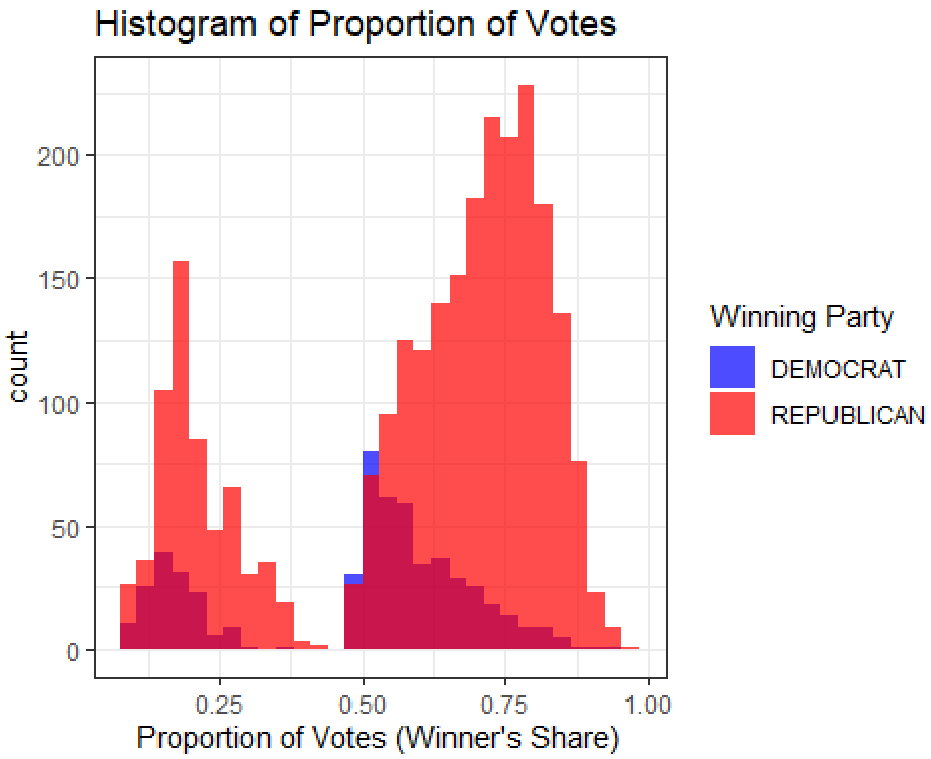
The histogram of weighted mean age of counties, categorized by the winning party, shows that distributions for both Democratic and Republican-won counties are centered around ages 38–40. There is a slight tendency for Republican-won counties to have slightly older populations, while Democratic-won counties tend to have younger populations.



- *Insights:* Age appears to be relatively normally distributed across counties, with subtle differences based on the winning party. This suggests that mean age might serve as a modest predictor of voting outcomes, potentially interacting with other demographic variables. Younger populations may favor Democrats, while older populations may lean Republican.

3.3 Proportion of Votes for the Winning Party

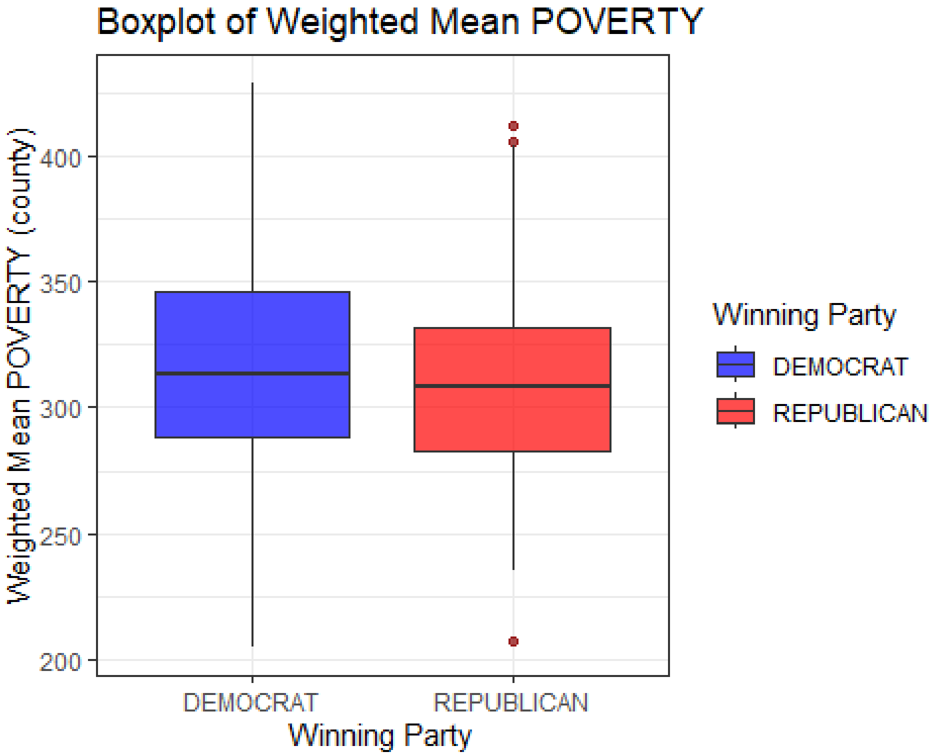
This histogram displays the distribution of the proportion of votes received by the winning party in each county. A clear distinction emerges: Republican-won counties (red) frequently exhibit higher vote proportions, often 75% or more. In contrast, Democratic-won counties (blue) tend to have closer margins of victory, with vote proportions often clustering around 50–60%.



- *Insights:* This pattern highlights differences in electoral dominance. Republican victories in counties are often more decisive, whereas Democratic wins tend to occur in more competitive counties. This could reflect underlying differences in voter concentration, engagement, or political homogeneity within counties.

3.4 Boxplot of Weighted Mean Poverty by Winning Party

The boxplot comparing weighted mean poverty levels in counties by the winning party indicates that poverty levels are slightly higher in counties won by Republicans, although the median poverty levels are relatively similar for both parties. The distribution of poverty levels is narrower and more consistent for Democratic-won counties. Republican-won counties, however, display greater variability in poverty levels, including several outliers, particularly at the higher end of the poverty spectrum.



- *Insights:* The variation in poverty levels suggests that economic conditions may influence voting patterns. The wider range of poverty levels in Republican-won counties (encompassing both very high- and low-poverty areas) compared to the more consistent levels in Democratic counties is noteworthy. Outliers, especially high-poverty Republican counties, may require specific attention to understand their role in electoral outcomes.

4. Modeling County-Level Vote Share (Linear Regression)

Following the exploratory analysis, linear regression models were developed to explain the variation in a continuous outcome variable: the county-level vote share for the winning presidential candidate in 2020 (`prop_winner_2020`). This response variable, expressed as a percentage, allows for an interpretation of electoral support independent of vote volume and helps stabilize variance. Predictor variables included quantitative and categorical demographic factors informed by theory and EDA findings, such as mean age, proportion of males, proportion with a 4-year college education, proportion white, and mean family income.

4.1 Full Linear Model

A multiple linear regression model was fitted using the selected demographic predictors.

Model Equation:

$$\widehat{\text{prop\_winner\_2020}} = -0.857 + 0.00898 \cdot \text{mean\_age} + 1.791 \cdot \text{prop\_male} + 0.401 \cdot \text{prop\_educ\_4yr} + 0.0350 \cdot \text{prop\_white} + 5.57 \times 10^{-8} \cdot \text{mean\_ftotinc}$$

Table 4.11: Coefficient Estimates for Predicting Winner Vote Share

term	estimate	std.error	statistic	p.value
(Intercept)	-0.857247	0.427320	-2.006100	4.5e-02
mean_age	0.008975	0.003273	2.742032	6.4e-03
prop_male	1.791495	0.797727	2.245749	2.5e-02
prop_educ_4yr	0.400591	0.214533	1.867272	6.3e-02
prop_white	0.034962	0.073285	0.477072	6.3e-01
mean_ftotinc	0.000000	0.000000	1.934313	5.4e-02

Table 4.12: Model Fit Statistics

r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual
0.0420	0.0307	0.19	3.72	2.6e-03	5	424

Interpretation of Results:

The linear model estimates the proportion of votes received by the winning presidential candidate in each county. Key findings include:

- **Proportion of Male Residents (`prop_male`):** This variable has a statistically significant and substantial positive coefficient (1.791, p = 0.025). This suggests that in counties with a higher proportion of male residents, the winning candidate tended to secure a larger share of the votes.

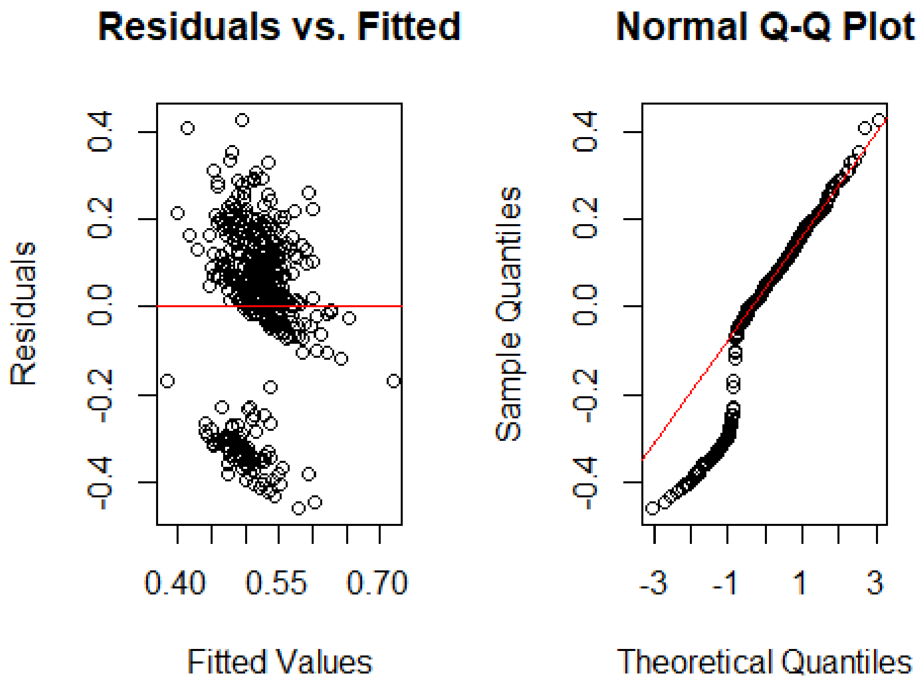
- **Mean Age (`mean_age`):** Mean age also demonstrates a positive and statistically significant relationship (0.00898, p = 0.006), indicating that counties with older populations were associated with a higher vote share for the winning candidate.

- **Proportion of College-Educated Residents (prop_educ_4yr) and Mean Family Income (mean_ftotinc):** These variables show marginal statistical significance (p = 0.063 and p = 0.054, respectively). Their positive coefficients suggest a slight tendency for the winner's vote share to increase as county-level education and income levels rise.
- **Proportion White (prop_white):** This variable was not statistically significant (p = 0.631), indicating no clear linear relationship between the proportion of white residents and the winning candidate's vote share in this model, after controlling for other factors.

Critically, the model possesses limited explanatory power. The **Adjusted R-squared value is 0.0307 (or 0.031)**, signifying that the included demographic variables explain only about 3.1% of the variability in the winner's vote share at the county level. This low R-squared value strongly suggests that while certain demographic indicators like gender composition and age are statistically relevant, a large portion of the variation in vote share is likely driven by other unmeasured factors. These could include political campaign intensity, the impact of local issues, pre-existing party affiliation strengths, voter turnout dynamics, or candidate-specific appeal.

4.2 Model Diagnostics

We run diagnostics of the model above.



Diagnostic plots for the linear model were examined:

- **Residuals vs. Fitted Plot:** This plot showed some clustering of points above and below the zero line across different fitted values, with a group of residuals below -0.2, hinting at potential heterogeneity. While no severe funnel shape indicative of glaring heteroscedasticity was observed, a slight curvature suggested that the simple linear form might not capture all underlying structures (e.g., interactions or non-linear terms).
- **Normal Q-Q Plot:** Residuals mostly aligned with the diagonal line, particularly between -2 and +1. However, deviations were noted in the tails (around -2 or below, and +3 or above), indicating mild departures from normality.

Overall, the assumptions of linearity, constant variance, and normality of errors appear to be moderately satisfied. There are indications that more complex modeling approaches, such as including interaction terms, non-linear transformations, or additional predictors, might improve the model's fit and predictive accuracy.

4.3 Reduced Model and Hypothese test

To assess the joint importance of `prop_educ_4yr` and `mean_ftotinc`, these variables were removed from the full model to create a reduced model. An ANOVA test was then conducted to compare the full and reduced models.

- **Null Hypothesis (H_0)**: The additional variables (`prop_educ_4yr` and `mean_ftotinc`) do not improve the model. In other words:

  $$\beta_{\text{prop\_educ\_4yr}} = \beta_{\text{mean\_ftotinc}} = 0$$

  This implies that the simpler model (Model 1) is sufficient to explain the variation in the dependent variable.

- **Alternative Hypothesis (H_1)**: At least one of the added predictors (`prop_educ_4yr` or `mean_ftotinc`) significantly improves the model. In other words:

  $$\beta_{\text{prop\_educ\_4yr}} \neq 0 \quad \text{or} \quad \beta_{\text{mean\_ftotinc}} \neq 0$$

Table 4.21: Coefficient Estimates (Reduced Model)

term	estimate	std.error	statistic	p.value
(Intercept)	-0.688478	0.405853	-1.696372	9.1e-02
mean_age	0.007451	0.003234	2.304011	2.2e-02
prop_male	1.800944	0.761814	2.364022	1.9e-02
prop_white	0.020887	0.073436	0.284419	7.8e-01

Table 4.22: ANOVA Model Comparison (Reduced vs Full)

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
426	15.3991				
424	15.1602	2	0.2389	3.3402	3.6e-02

The ANOVA results show a decrease in the Residual Sum of Squares (RSS) from 15.399 in the reduced model to 15.160 in the full model. The F-statistic for this comparison is 3.3402, with a p-value of 0.036. Since this p-value is less than the conventional significance level of α = 0.05, we reject the null hypothesis. This indicates that the inclusion of `prop_educ_4yr` and `mean_ftotinc` together significantly improves the model's ability to explain variation in the winner's vote share, suggesting that at least one of these predictors makes a meaningful contribution.

4.4 Confidence Intervals for a New Observation

To illustrate the model's predictive capability for a hypothetical county with specific demographic characteristics (mean_age = 45, prop_male = 0.48, prop_educ_4yr = 0.25, prop_white = 0.70, mean_ftotinc = 50000), confidence and prediction intervals were computed.

Table 4.41: Confidence Interval for Mean Response

fit	lwr	upr
0.5340	0.4777	0.5902

Table 4.42: Prediction Interval for New Observation

fit	lwr	upr
0.5340	0.1581	0.9099

- **Mean Response Confidence Interval (95% CI):** [0.4777, 0.5902], with a fitted value of 0.5340. This interval estimates the average vote share for the winning candidate in counties with these exact demographic characteristics.
- **New Observation Prediction Interval (95% CI):** [0.1581, 0.9099], with a fitted value of 0.5340. This interval predicts the vote share for a single, specific new county with these demographics. The much wider range reflects the additional uncertainty associated with predicting an individual outcome rather than an average.

5. Modeling Likelihood of Republican Win (Logistic Regression)

To further explore electoral dynamics, a logistic regression model was developed to predict the probability of a Republican candidate winning a county (`rep_win` = 1). This approach shifts the focus from the *share* of votes to the binary *outcome* of which party wins.

Logistic Regression Model Equation (Log-Odds Form):

$$\log\left(\frac{\mathbb{P}(\text{rep\_win} = 1)}{1 - \mathbb{P}(\text{rep\_win} = 1)}\right)$$
$$= -9.451 + 0.124 \cdot \text{mean\_age} + 0.303 \cdot \text{prop\_male} + 16.638 \cdot \text{prop\_white} + 9.782 \cdot \text{prop\_black}$$
$$- 14.155 \cdot \text{prop\_asian} + 6.519 \cdot \text{prop\_amind} - 27.923 \cdot \text{prop\_educ\_4yr} + 0.373 \cdot \text{log\_med\_inc}$$

Table 5.01: Logistic Regression Coefficient Estimates for Predicting Republican Win

term	estimate	std.error	statistic	conf.low	conf.high	p.value
(Intercept)	-9.4509	3.4818	-2.7144	-16.6282	-2.8831	6.6e-03
mean_age	0.1240	0.1371	0.9047	-0.1419	0.3976	3.7e-01
prop_male	0.3029	0.1686	1.7960	-0.0137	0.6527	7.2e-02
prop_white	16.6382	3.7917	4.3880	9.5087	24.4764	1.1e-05
prop_black	9.7821	3.7840	2.5851	2.5786	17.5183	9.7e-03
prop_asian	-14.1552	10.6347	-1.3310	-35.6504	6.1418	1.8e-01
prop_amind	6.5186	8.2973	0.7856	-16.3314	20.1451	4.3e-01
prop_educ_4yr	-27.9234	6.0014	-4.6528	-40.0345	-16.4473	3.3e-06
log_med_inc	0.3727	0.2347	1.5879	-0.0862	0.8375	1.1e-01

Table 5.02: Logistic Regression Fit Statistics

null.deviance	df.null	deviance	df.residual	AIC
595.77	429	360.26	421	378.26

Interpretation of Results:

This model identifies demographic and economic features associated with an increased or decreased likelihood of a Republican win at the county level.

- **Proportion White (`prop_white`):** This variable has a very large, positive, and statistically significant coefficient (16.638, p < 0.001). This suggests that counties with higher proportions of white residents are substantially more likely to vote Republican.

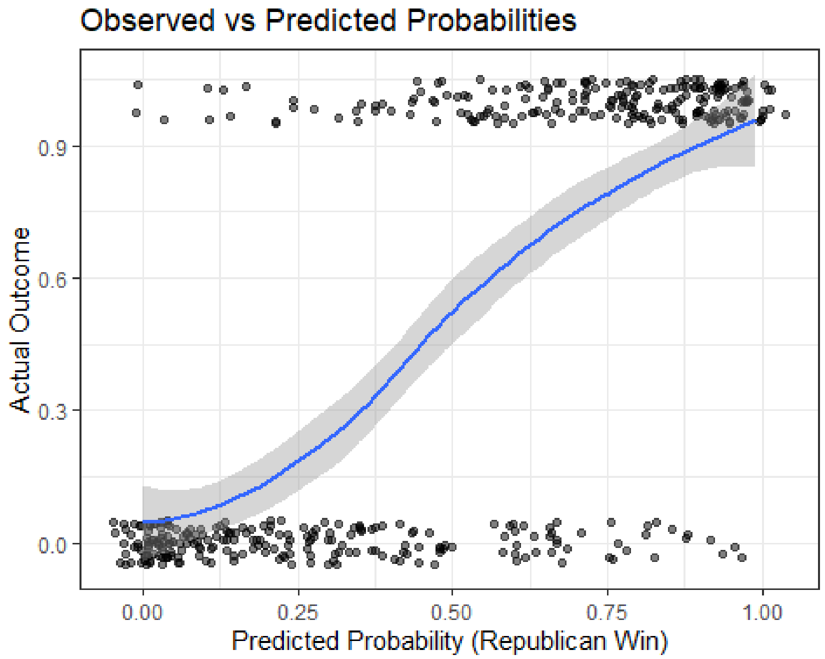
- **Proportion Black (`prop_black`):** This variable also shows a statistically significant positive association with Republican wins (9.782, p = 0.0097). This finding is counterintuitive to general national voting patterns of individuals and, as noted in the original draft, may reflect multicollinearity with other racial composition variables (especially `prop_white`) or act as a suppressor variable within this specific model structure and county-level unit of analysis. It does not imply that Black individuals in those counties voted Republican at higher rates, but rather that counties with these particular demographic compositions (when all variables are considered) showed this association.

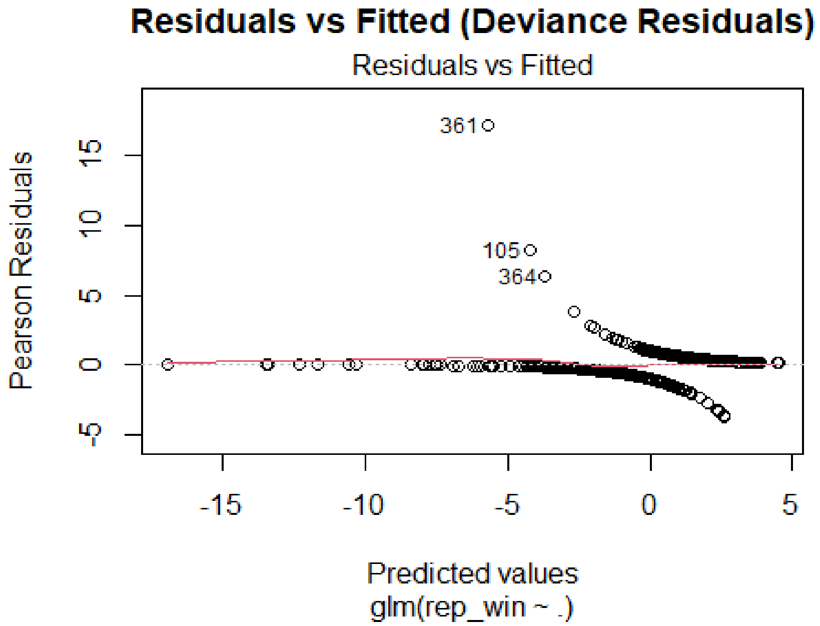
- **Proportion with a 4-Year Degree (`prop_educ_4yr`):** This variable is strongly and negatively associated with Republican wins (–27.923, p < 0.001). This implies that counties with a more highly educated populace (specifically, a higher proportion of college graduates) are significantly less likely to vote Republican.

- **Proportion Male (`prop_male`):** Shows marginal significance (p = 0.072) with a positive coefficient, suggesting a weak tendency for counties with more males to favor Republican candidates.

- **Other Variables:** `mean_age`, `prop_asian`, `prop_amind`, and `log_med_inc` did not emerge as statistically significant predictors of Republican victory in this particular model specification.

5.1 Logistic Regression Diagnostics





- **Observed vs. Predicted Probabilities Plot:** This plot demonstrated good model calibration. The S-shaped loess curve, representing the average trend across predicted probabilities, smoothly transitioned from 0 to 1, with most actual outcomes aligning with the predicted probability range. This confirms that higher predicted probabilities of a Republican win generally corresponded with actual Republican wins.

- **Residuals vs. Fitted Plot (Deviance Residuals):** While logistic regression does not assume homoscedasticity or linear residuals, deviance residual plots still help flag model misspecification. Most residuals clustered around zero. A few high-leverage outliers were identified, which might warrant further investigation. However, no strong pattern in residual spread suggested severe violations of model assumptions. Slight curvature could indicate potential non-linear relationships not captured by the current model.

- **Multicollinearity:** Given the inclusion of several race proportion variables (`prop_white`, `prop_black`, `prop_asian`, `prop_amind`), multicollinearity is a concern that could affect the stability and interpretation of individual coefficients, particularly for these variables.

5.2 Confidence Intervals for Logistic Regression Coefficients

Table 5.2: 95% Confidence Intervals for Logistic Regression Coefficients

Term	95% CI (Low)	95% CI (High)
(Intercept)	-16.6282	-2.8831
mean_age	-0.1419	0.3976
prop_male	-0.0137	0.6527
prop_white	9.5087	24.4764
prop_black	2.5786	17.5183
prop_asian	-35.6504	6.1418
prop_amind	-16.3314	20.1451
prop_educ_4yr	-40.0345	-16.4473
log_med_inc	-0.0862	0.8375

The 95% confidence intervals reinforce the significance findings:

`prop_white`: [9.51, 24.48] (excludes zero)

`prop_black`: [2.58, 17.52] (excludes zero)

`prop_educ_4yr`: [–40.03, –16.45] (excludes zero)

Intervals for `mean_age`, `prop_male` (marginally, as it’s close to including zero), `prop_asian`, `prop_amind`, and `log_med_inc` all include zero, indicating statistical uncertainty about their effects in this model.

6. Discussion

This project aimed to elucidate the influence of demographic factors on U.S. presidential election outcomes at the county level, using the 2020 election as a case study. The findings from both the linear regression (predicting the winner's vote share) and logistic regression (predicting Republican party wins) models offer several key insights while also highlighting the inherent complexities of electoral analysis.

Synthesis of Findings:

The linear regression model for the winner's vote share revealed that counties with older mean ages and higher proportions of male residents tended to give a larger vote share to the eventual winning candidate. The joint significance of education and income (from the ANOVA test) suggests these factors also play a role, though their individual p-values in the full model were marginal. However, the most striking aspect of this model was its low Adjusted R-squared value (0.031). This indicates that the selected demographic variables, while some are statistically significant, explain a very small fraction of the county-level variation in the winning candidate's vote share. This aligns with perspectives suggesting that demographics alone are insufficient to fully predict or explain electoral outcomes, pointing towards the importance of other unmodeled factors like campaign strategies, local political contexts, candidate-specific appeal, and voter turnout initiatives.

The logistic regression model, predicting the likelihood of a Republican win, provided a different lens. Here, racial composition and education level emerged as powerful predictors. A higher proportion of white residents (`prop_white`) and, counterintuitively, a higher proportion of Black residents (`prop_black`) were strongly associated with an increased probability of a Republican county win. Conversely, a higher proportion of residents with a four-year college degree (`prop_educ_4yr`) was strongly associated with a decreased probability of a Republican win. The positive coefficient for `prop_black` warrants careful interpretation; it is likely an artifact of the county-level aggregation and potential multicollinearity with other racial demographic variables or suppressor effects, rather than a reflection of individual voting behavior. It underscores that ecological inferences (drawing conclusions about individuals from aggregated data) must be made with extreme caution.

Contrasting Insights:

It's noteworthy that `prop_white` was not a significant predictor of the *winner's vote share* in the linear model but was a highly significant predictor of a *Republican win* in the logistic model. Similarly, `prop_educ_4yr` had a marginal effect on the winner's vote share but a very strong negative effect on the likelihood of a Republican win. These differences underscore how the choice of outcome variable (continuous vote share vs. binary party win) and model specification can reveal different facets of demographic influence.

Limitations:

This study is subject to several limitations:

1. **Ecological Fallacy:** The analysis is at the county level. Relationships observed at this aggregated level do not necessarily translate to individual voter behavior.

2. **Omitted Variable Bias:** The low R-squared in the linear model explicitly points to the influence of unmeasured variables. Factors such as party identification, political ideology, campaign spending, local media environment, and specific policy issues are not included but are known to affect election results.

3. **Multicollinearity:** Particularly in the logistic regression with multiple racial proportion variables, multicollinearity may affect the precision and interpretation of individual coefficient estimates. The counterintuitive sign for `prop_black` could be a symptom of this.

4. **Cross-Sectional Data:** The primary analysis focuses on the 2020 election, limiting the ability to draw conclusions about dynamic changes over time without further longitudinal modeling.

5. **Definition of "Winner":** The linear model predicts the "winner's" vote share, which could be a Democrat or a Republican. This might obscure party-specific demographic effects if the demographic drivers for winning differ significantly between parties.

7. Conclusion and Future Work

This paper demonstrates a methodological approach to modeling U.S. election outcomes using demographic and census data at the county level. The findings indicate that while certain demographic factors like age, sex composition, racial makeup, and education levels exhibit statistically significant associations with voting patterns and party success, their collective ability to explain the variance in election outcomes (specifically, the winner's vote share) is limited in the linear model presented. This underscores the multifaceted nature of electoral politics, where demographics are one piece of a larger puzzle. The logistic regression provided clearer distinctions, particularly regarding the role of racial composition and education in predicting Republican party wins, though some findings necessitate cautious interpretation due to potential statistical artifacts like multicollinearity.

The results contribute to the ongoing discussion about the predictive power of demographics, suggesting that while they offer valuable insights, they are not solely deterministic. The low explanatory power of the linear model, in particular, highlights that a significant portion of electoral behavior is influenced by factors beyond the demographic variables considered here.

Future research could advance this work in several directions:

1. **Incorporating Additional Variables:** Future models could benefit from including variables related to local economic conditions beyond income (e.g., unemployment change, industry composition), measures of political ideology or partisanship at the county level, campaign spending data, and voter turnout rates.

2. **Exploring Interactions and Non-Linearities:** The diagnostic plots suggested potential non-linear relationships. Future work could explore interaction terms between demographic variables (e.g., age and education) and non-linear transformations of predictors.

3. **Longitudinal Analysis:** Expanding the analysis across multiple election cycles (using the 2010-2020 data more dynamically) could reveal trends in how the influence of demographic factors has changed over time.

4. **Geographically Weighted Regression (GWR):** Given the spatial nature of election data, GWR could be employed to explore how relationships between demographics and voting patterns vary across different regions of the U.S.

5. **Advanced Machine Learning Techniques:** Employing machine learning models (e.g., random forests, gradient boosting) might capture more complex relationships and improve predictive accuracy, though often at the cost of direct interpretability of coefficients.

6. **Addressing Multicollinearity:** Further investigation into multicollinearity, perhaps through techniques like Principal Component Analysis (PCA) for demographic clusters or by modeling racial proportions differently, could yield more stable and interpretable coefficients for related variables.

7. **Simpler Relationship Exploration:** As initially noted in the draft, delving into simpler, bivariate relationships or more focused multivariate models with fewer predictors might sometimes yield clearer, more actionable insights for specific demographic segments.

In conclusion, while the current models provide a foundational understanding, the path to more comprehensively explaining and predicting election outcomes through demographic data requires continued refinement, the integration of a broader array of influencing factors, and the application of diverse analytical techniques.

Appendix

Below is the code we used to load, summarize, visualize and model our data.

```r
library(readr)
library(data.table)
library(dplyr)
library(ggplot2)
library(stringr)
library(forcats)
library(sf)
library(ggpubr)
library(broom)
library(car)
library(broom)
library(flextable)
library(knitr)
library(ResourceSelection)
library(ROCR)


# Read data
election_data <- read_csv("ProjectData/countypres_2000-2020.csv")
population_data <- read_csv("ProjectData/Population.csv")
diploma_higher_data <- read_csv("ProjectData/Percent of Diploma or Higher - 1940 - 2000.csv")
census_data <- fread("ProjectData/usa_00004.csv")

#str(election_data)
#str(population_data)
#str(diploma_higher_data)
#str(census_data)

election_2020 <- election_data %>%
  filter(year == 2020 & office == "US PRESIDENT") %>%
  group_by(state, state_po, county_name, county_fips, party) %>%
  summarise(total_votes_2020 = sum(totalvotes, na.rm = TRUE),
            candidate_votes = sum(candidatevotes, na.rm = TRUE),
            .groups = "drop") %>%
  group_by(state, state_po, county_name, county_fips) %>%
  mutate(prop_votes = candidate_votes / total_votes_2020) %>%
  ungroup()


winners_2020 <- election_2020 %>%
  group_by(county_fips) %>%
  filter(prop_votes == max(prop_votes, na.rm = TRUE)) %>%
  ungroup() %>%
  select(county_fips, party, prop_votes) %>%
  rename(party_winner_2020 = party,
         prop_winner_2020 = prop_votes)


acs_by_county <- census_data %>%

  mutate(COUNTYFIP = paste0(STATEFIP, sprintf("%03d", COUNTYFIP))) %>%
  group_by(COUNTYFIP) %>%
  summarise(

    mean_FTOTINC = weighted.mean(FTOTINC, w=PERWT, na.rm=TRUE),


    mean_AGE = weighted.mean(AGE, w=PERWT, na.rm=TRUE),


    prop_male = weighted.mean(SEX == 1, w=PERWT, na.rm=TRUE),
    prop_female = weighted.mean(SEX == 2, w=PERWT, na.rm=TRUE),


    mean_POVERTY = weighted.mean(POVERTY, w=PERWT, na.rm=TRUE),


    prop_college = weighted.mean(EDUCD >= 65, w=PERWT, na.rm=TRUE)
  ) %>%
  ungroup()



analysis_df <- winners_2020 %>%
  mutate(county_fips = as.character(county_fips)) %>%  # Convert to character
  left_join(acs_by_county, by = c("county_fips" = "COUNTYFIP")) %>%
  rename(
    fips = county_fips
  )

ggplot(analysis_df, aes(x = mean_FTOTINC, fill = party_winner_2020)) +
  geom_histogram(bins = 30, alpha = 0.7, position = "identity") +
  theme_bw() +
  scale_fill_manual(values = c("DEMOCRAT" = "blue", "REPUBLICAN" = "red")) +
  labs(title = "Distribution of County-Level Mean Family",
       x = "Weighted Mean Family Income (County)",
       fill = "Winning Party")

# Histogram for Mean Age
ggplot(analysis_df, aes(x = mean_AGE, fill = party_winner_2020)) +
  geom_histogram(bins = 30, alpha = 0.7, position = "identity") +
  theme_bw() +
  scale_fill_manual(values = c("DEMOCRAT" = "blue", "REPUBLICAN" = "red")) +
  labs(title = "Distribution of County-Level Mean AGE",
       x = "Weighted Mean Age (County)",
       fill = "Winning Party")

ggplot(analysis_df, aes(x = prop_winner_2020, fill = party_winner_2020)) +
  geom_histogram(bins = 30, alpha = 0.7, position = "identity") +
  theme_bw() +
  scale_fill_manual(values = c("DEMOCRAT" = "blue", "REPUBLICAN" = "red")) +
  labs(
    title = "Histogram of Proportion of Votes",
    x = "Proportion of Votes (Winner's Share)",
    fill = "Winning Party"
  )

ggplot(analysis_df, aes(x = party_winner_2020, y = mean_POVERTY, fill =
party_winner_2020)) +
  geom_boxplot(alpha = 0.7, outlier.color = "darkred") +
  theme_bw() +
  scale_fill_manual(values = c("DEMOCRAT" = "blue", "REPUBLICAN" = "red",
"OTHER" = "purple")) +
  labs(
    title = "Boxplot of Weighted Mean POVERTY",
    x = "Winning Party",
    y = "Weighted Mean POVERTY (county)",
    fill = "Winning Party"
  )

library(dplyr)

census_agg <- census_data %>%
  # Filter to a single year (e.g., 2020) if it makes sense
  filter(YEAR == 2020) %>%

  mutate(COUNTYFIP = paste0(STATEFIP, sprintf("%03d", COUNTYFIP))) %>%
  group_by(COUNTYFIP) %>%
  summarise(
    # -- Age & Sex --
    mean_age         = mean(AGE, na.rm = TRUE),
    prop_male        = mean(SEX == 1, na.rm = TRUE),
    prop_female      = mean(SEX == 2, na.rm = TRUE),

    # -- Race Proportions (pick whichever categories you plan to analyze) --
    prop_white       = mean(RACE == 1, na.rm = TRUE),
    prop_black       = mean(RACE == 2, na.rm = TRUE),
    prop_amind       = mean(RACE == 3, na.rm = TRUE),  # American
Indian/Alaska Native
    prop_chinese     = mean(RACE == 4, na.rm = TRUE),
    prop_japanese    = mean(RACE == 5, na.rm = TRUE),
    prop_other_asian = mean(RACE == 6, na.rm = TRUE),  # "Other Asian or
```

```
Pacific Islander"
    prop_other_races = mean(RACE %in% c(7, 8, 9), na.rm = TRUE),

    prop_educ_4yr    = mean(EDUC == 10, na.rm = TRUE),

    mean_ftotinc     = mean(FTOTINC, na.rm = TRUE),
    median_ftotinc   = median(FTOTINC, na.rm = TRUE)
  ) %>%
  ungroup()

analysis_data <- winners_2020 %>%
  mutate(COUNTYFIP = as.character(county_fips)) %>%  # Convert to character
  left_join(census_agg, by = "COUNTYFIP") %>%
  filter(
    !is.na(prop_winner_2020),
    !is.na(mean_age),
    !is.na(prop_white),
    !is.na(mean_ftotinc)  # or median_ftotinc
  )

# Fit model
lm_full <- lm(
  prop_winner_2020 ~ mean_age + prop_male + prop_educ_4yr + prop_white +
mean_ftotinc,
  data = analysis_data
)

#summary(lm_full)

model_summary <- tidy(lm_full) %>%
  mutate(
    p.value_fmt = ifelse(p.value < 2e-16, "< 2e-16", format(p.value,
scientific = TRUE, digits = 2))
  ) %>%
  select(term, estimate, std.error, statistic, p.value_fmt) # <-- select
BEFORE flextable

coef_table <- flextable(model_summary) %>%
  set_caption("Table 4.11: Coefficient Estimates for Predicting Winner Vote
Share") %>%
  colformat_double(j = c("estimate", "std.error", "statistic"), digits = 6)
%>%
  italic(j = 1, part = "header") %>%
  set_header_labels(p.value_fmt = "p.value") %>%
  autofit()

# Model stats with formatted p-values
model_stats <- glance(lm_full) %>%
  mutate(
    p.value_fmt = ifelse(p.value < 2e-16, "< 2e-16", format(p.value,
scientific = TRUE, digits = 2))
  ) %>%
  select(r.squared, adj.r.squared, sigma, statistic, p.value_fmt, df,
df.residual) # <-- select BEFORE flextable

model_stats_table <- flextable(model_stats) %>%
  set_caption("Table 4.12: Model Fit Statistics") %>%
  colformat_double(j = c("r.squared", "adj.r.squared"), digits = 4) %>%
  colformat_double(j = c("sigma", "statistic"), digits = 2) %>%
  set_header_labels(p.value_fmt = "p.value") %>%
  colformat_double(j = c("df", "df.residual"), digits = 0) %>%
  autofit()

coef_table
model_stats_table

# Plot residual diagnostics (conceptual; results go to Appendix)
par(mfrow = c(1, 2))
plot(lm_full$fitted.values, lm_full$residuals,
     xlab = "Fitted Values",
     ylab = "Residuals",
     main = "Residuals vs. Fitted")
abline(h=0, col="red")

qqnorm(lm_full$residuals, main="Normal Q-Q Plot")
qqline(lm_full$residuals, col="red")
par(mfrow = c(1, 1))

lm_reduced <- lm(
  prop_winner_2020 ~ mean_age + prop_male + prop_white,
  data = analysis_data
)


reduced_summary <- tidy(lm_reduced) %>%
  mutate(
    p.value_fmt = ifelse(p.value < 2e-16,
                         "< 2e-16",
                         format(p.value, scientific = TRUE, digits = 2))
  ) %>%
  select(term, estimate, std.error, statistic, p.value_fmt)

coef_table_reduced <- flextable(reduced_summary) %>%
  set_caption("Table 4.21: Coefficient Estimates (Reduced Model)") %>%
  colformat_double(j = c("estimate", "std.error", "statistic"), digits = 6)
%>%
  italic(j = 1, part = "header") %>%
  set_header_labels(p.value_fmt = "p.value") %>%
```

```r
  autofit()


anova_result <- anova(lm_reduced, lm_full)
anova_df <- as.data.frame(anova_result)


if ("Pr(>F)" %in% names(anova_df)) {
  anova_df <- anova_df %>%
    mutate(
      PrF_fmt = ifelse(`Pr(>F)` < 2e-16,
                       "< 2e-16",
                       format(`Pr(>F)`, scientific = TRUE, digits = 2))
    )
} else {
  anova_df$PrF_fmt <- NA_character_
}

anova_table <- flextable(anova_df %>%
  mutate(across(where(is.numeric), ~round(., 4))) %>% # optional: round
numeric columns
  select(Res.Df, RSS, Df, `Sum of Sq`, F, PrF_fmt)) %>%
  set_caption("Table 4.22: ANOVA Model Comparison (Reduced vs Full)") %>%
  set_header_labels(PrF_fmt = "Pr(>F)") %>%
  autofit()


coef_table_reduced
anova_table

# Full model
new_county <- data.frame(
  mean_age = 45,
  prop_male = 0.48,
  prop_educ_4yr = 0.25,
  prop_white = 0.70,
  mean_ftotinc = 50000
)

conf_pred <- predict(lm_full, newdata = new_county, interval = "confidence")
pred_pred <- predict(lm_full, newdata = new_county, interval = "prediction")

conf_df <- as.data.frame(conf_pred)
pred_df <- as.data.frame(pred_pred)

conf_table <- flextable(conf_df) %>%
  set_caption("Table 4.41: Confidence Interval for Mean Response") %>%
  colformat_double(digits = 4) %>%
  autofit()
```

```r
pred_table <- flextable(pred_df) %>%
  set_caption("Table 4.42: Prediction Interval for New Observation") %>%
  colformat_double(digits = 4) %>%
  autofit()

conf_table
pred_table

dat <- analysis_data %>%
  mutate(
    rep_win = as.integer(party_winner_2020 == "REPUBLICAN"),
    # collapse Asian sub-groups
    prop_asian = prop_chinese + prop_japanese + prop_other_asian,
    log_med_inc = scale(log(median_ftotinc), scale = TRUE, center =
TRUE)[,1],
    mean_age    = scale(mean_age, scale = TRUE, center = TRUE)[,1],
    prop_male   = scale(prop_male, scale = TRUE, center = TRUE)[,1]
  ) %>%
  # drop “other races” to break the linear dependence
  select(rep_win, mean_age, prop_male, prop_white, prop_black,
         prop_asian, prop_amind, prop_educ_4yr, log_med_inc)

glm_clean <- glm(rep_win ~ ., data = dat, family = binomial)


glm_summary <- tidy(glm_clean, conf.int = TRUE) %>%
  mutate(
    p.value_fmt = ifelse(p.value < 2e-16,
                         "< 2e-16",
                         format(p.value, scientific = TRUE, digits = 2))
  ) %>%
  select(term, estimate, std.error, statistic, conf.low, conf.high,
p.value_fmt)

glm_coef_table <- flextable(glm_summary) %>%
  set_caption("Table 5.01: Logistic Regression Coefficient Estimates for
Predicting Republican Win") %>%
  colformat_double(j = c("estimate", "std.error", "statistic", "conf.low",
"conf.high"), digits = 4) %>%
  italic(j = 1, part = "header") %>%
  set_header_labels(p.value_fmt = "p.value") %>%
  autofit()

# Assuming AIC is uppercase
glm_stats <- glance(glm_clean) %>%
  select(null.deviance, df.null, deviance, df.residual, AIC)

glm_stats_table <- flextable(glm_stats) %>%
  set_caption("Table 5.02: Logistic Regression Fit Statistics") %>%
  colformat_double(digits = 2) %>%
  autofit()


#summary(glm_clean)
glm_coef_table
glm_stats_table

dat$predicted_prob <- predict(glm_clean, type = "response")
ggplot(dat, aes(x = predicted_prob, y = rep_win)) +
  geom_jitter(width = 0.05, height = 0.05, alpha = 0.5) +
  geom_smooth(method = "loess", formula = y ~ x) +
  labs(title = "Observed vs Predicted Probabilities", x = "Predicted
Probability (Republican Win)", y = "Actual Outcome") +
  theme_bw()


plot(glm_clean, which = 1, main = "Residuals vs Fitted (Deviance Residuals)")

# 5. Inference: Confidence Intervals for Coefficients
confint_glm <- confint(glm_clean) # default is 95%
confint_glm <- as.data.frame(confint_glm)
names(confint_glm) <- c("95% CI (Low)", "95% CI (High)")
confint_glm <- tibble::rownames_to_column(confint_glm, "Term")

confint_table <- flextable(confint_glm) %>%
  set_caption("Table 5.2: 95% Confidence Intervals for Logistic Regression
Coefficients") %>%
  colformat_double(j = c("95% CI (Low)", "95% CI (High)"), digits = 4) %>%
  autofit()

confint_table
```