My title*

My subtitle if needed

Xizi Sun

November 25, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Data Tool

The dataset was analyzed using R (R Core Team 2023) and utilized several R packages for data manipulation and visualization, including ggplot2 (Wickham 2016), dplyr (Wickham et al. 2023), and here (Müller and Bryan 2023). The data was processed using the Apache Arrow package (Richardson, Dunnington, and Developers 2023) for efficient file handling and visualized with scales (Wickham, Seidel, et al. 2023) for enhanced graphics. Reproducibility was ensured using knitr (Xie 2014). Data was extracted from IPUMS USA (IPUMS 2024). Guidance on storytelling with data was drawn from Telling Stories with Data (Alexander 2023).

*Code and data are available at: https://github.com/LilianS77/US_Marriage.

2.2 Data Source And Measurement

The data used for this study was sourced from **IPUMS USA (Integrated Public Use Microdata Series)** (IPUMS 2024), a comprehensive repository providing access to harmonized census and survey data from the United States. IPUMS USA is renowned for its meticulous process of standardizing variables across datasets, ensuring compatibility over time, and providing extensive metadata to support research. This harmonization process enables researchers to conduct longitudinal and cross-sectional studies on social, demographic, and economic trends.

The dataset consists of **individual-level microdata**, where each record represents a single person, numerically coded for all relevant characteristics. These records are organized into households, allowing for the study of individual behaviors and characteristics in the context of their family or co-residential settings. Unlike compiled statistics or pre-aggregated tables, this microdata structure provides researchers with unparalleled flexibility in exploring relationships between variables.

To address the diversity of record layouts, coding schemes, and documentation across the historical scope of the dataset, IPUMS implements a rigorous **harmonization process**. Variables are assigned **uniform codes**, ensuring consistency across census years (1850–2010) and the American Community Surveys (ACS) (2000–present). This standardization simplifies the analysis of long-term trends and facilitates comparisons across time and space.

2.3 Variable Selection

Using the IPUMS data extraction system, this study selected a focused subset of variables to examine the determinants of **non-marriage** in the United States. These include key demographic and socioeconomic characteristics, which are numerically coded for statistical analysis.

The table Table 1 shows the variables after data cleaning,

Table 1: Sample of cleaned data

Marital Status	Age	Gender	Race	Income	Education Level
Not_Married	73	Male	White	33900	Above_Bachelor
Married	43	Female	Asian	40000	High_School
Married	80	Female	White	13000	High_School
Married	66	Male	White	48000	Some_College
Married	52	Female	White	38100	Bachelor
Married	45	Female	American Indian	65000	High_School

2.3.1 Outcome variables

The primary outcome variable for this study is **Marital Status**, which categorizes individuals based on their marital state. The proportion of marital status categories is displayed in figure Figure 1. This variable allows for a comparison between individuals who have never married (`Not_Married`) and those who have (`Married`). For this study, **Not_Married**: Includes individuals who have never been married. **Married**: Includes individuals who are married, as well as those who are divorced, widowed, or separated.

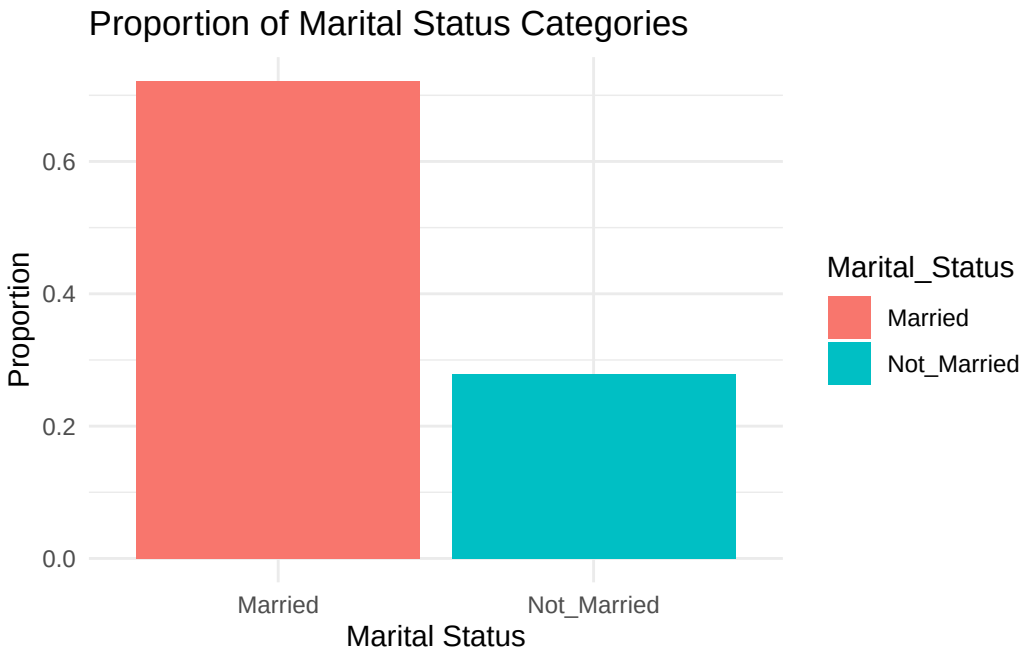


Figure 1: Proportion of Marital Status Categories

2.3.2 Predictor variables

This section focuses on the predictor variables included in the study. The distribution of predictor variables is displayed in Figure Figure 2. These variables capture demographic, socioeconomic, and personal characteristics, providing a comprehensive framework for analyzing factors associated with marital status. Below are the key predictor variables:

1. **Age**: A continuous variable representing the respondent's age in years.
2. **Gender**: A categorical variable indicating whether the respondent is male or female. Gender differences often play a role in marital patterns.

3. **Race**: A categorical variable categorized into White, Black, Asian, American Indian, and Other racial groups. This variable examines potential racial disparities in marital behavior.
4. **Education Level**: An ordinal variable indicating the highest level of education attained by the respondent. It is grouped into five categories: Below High School, High School, Some College, Bachelor's Degree, and Above Bachelor.
5. **Income**: A continuous variable measuring the respondent's annual income in dollars. Income reflects economic resources and may be associated with marital stability and decisions.

2.4 Data Selection

To understand the phenomenon of non-marriage in the United States, I selected the IPUMS USA dataset over alternatives such as IPUMS International. While IPUMS International provides harmonized census data from 104 countries with over 1 billion person records, its vast scope makes it less targeted for this study. My focus is specifically on the U.S. population, and IPUMS USA offers high-precision data from American Community Surveys (ACS) and federal censuses, making it a more suitable choice for studying societal patterns specific to the United States.

3 Model

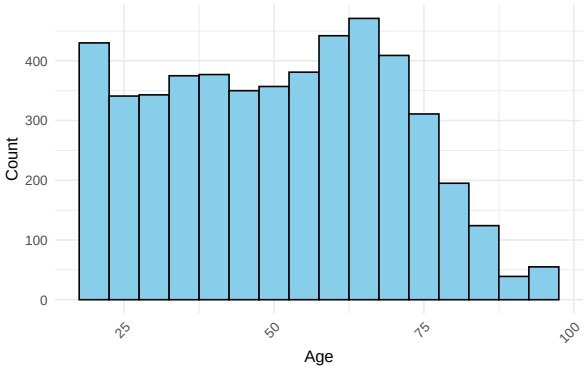
For this analysis, I employed a Logistic Regression Model to assess the likelihood of an individual not being married (outcome variable) based on several demographic and socioeconomic predictors (predictor variables). This model was chosen due to the binary nature of the outcome variable, which distinguishes between individuals who are "Not Married" versus those who are married (including divorced, widowed, or separated)..

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in **?@sec-model-details**.

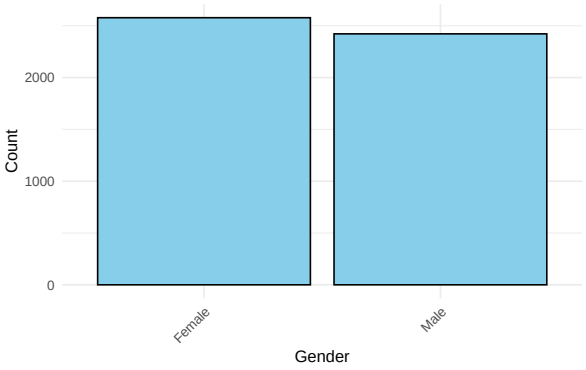
3.1 Model Setup

3.1.1 Objective

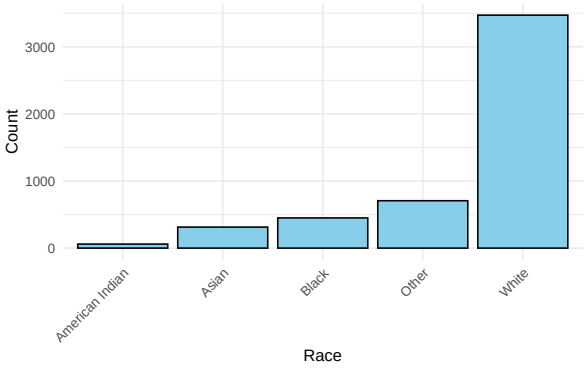
The primary objective of the model is to analyze and predict the factors associated with individuals' marital status, focusing on identifying key predictors for individuals who have never been married (`Not_Married`).



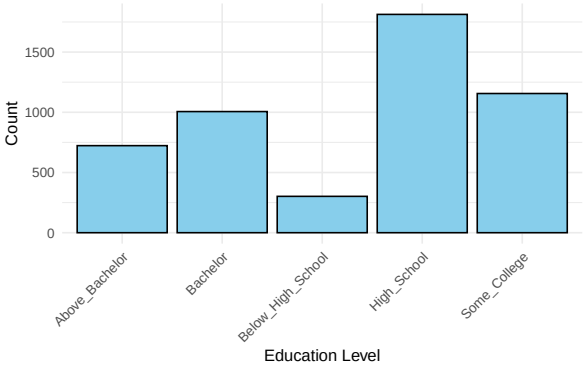
(a) Age Distribution



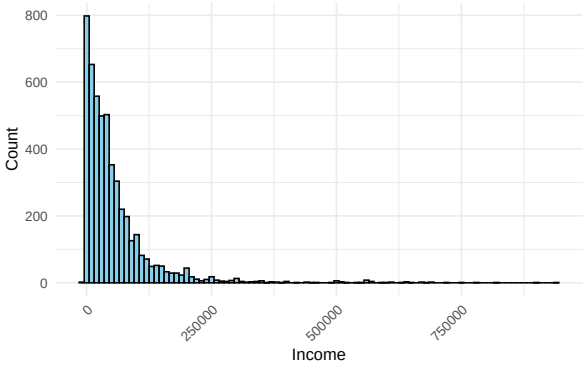
(b) Gender Distribution



(c) Race Distribution



(d) Education Level Distribution



(e) Income Distribution

Figure 2: fig-predictor-variables

3.1.2 Model Type

A **logistic regression model** is chosen for this analysis, as the outcome variable (`marital_status`) is binary (e.g., `Not_Married` vs. `Married`).

3.1.3 Outcome Variable

- `marital_status`: Encoded as:
  - `1` for `Not_Married`
  - `0` for `Married`

3.1.4 Predictor Variables

The predictors include: 1. **Age (`age`)**: A continuous variable representing the individual's age. 2. **Gender (`gender`)**: A categorical variable with levels `Male` and `Female`. 3. **Race (`Race`)**: A categorical variable with multiple racial categories (e.g., `White`, `Black`, etc.). 4. **Income (`Income`)**: A continuous variable representing the individual's annual income. 5. **Education Level (`education_level`)**: An ordinal variable with levels `Below_High_School`, `High_School`, `Some_College`, `Bachelor`, and `Above_Bachelor`.

3.1.5 Model Assumptions

1. **Linearity**: The log odds of the outcome are linearly related to the predictors.
2. **Independence**: Observations are independent of each other.
3. **No Multicollinearity**: Predictors are not highly correlated.
4. **No Outliers or Influential Points**: Checked through residual analysis.

3.1.6 Mathematical Representation

$$\log\left(\frac{P(\text{Not\_Married})}{1-P(\text{Not\_Married})}\right) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{Race} + \beta_4 \cdot \text{Income} + \beta_5 \cdot \text{education\_level}$$

Where: - (P(Not_Married)) is the probability of being `Not_Married`.

3.1.7 Model Implementation Steps

1. **Data Preprocessing**
   - Encode categorical variables (e.g., `gender`, `Race`, `education_level`) using one-hot encoding or dummy variables.
   - Normalize continuous variables like `age` and `Income` for better model performance.
   - Split the data into training and testing sets (e.g., 80% train, 20% test).
2. **Model Building**
   - Use the `glm` function in R for logistic regression modeling.
   - Define the formula as:

```
glm(marital_status ~ age + gender + Race + Income + education_level,
    data = training_data,
    family = "binomial")
```

3. **Validation and Performance**
   - Evaluate the model using metrics such as:
     - **Accuracy**: Proportion of correctly predicted observations.
     - **ROC Curve and AUC**: To assess the tradeoff between sensitivity and specificity.
     - **Confusion Matrix**: To evaluate true positives, true negatives, false positives, and false negatives.
4. **Interpretation**
   - Analyze the coefficients (( )) to determine the relationship between predictors and the likelihood of being `Not_Married`.

3.1.8 Potential Limitations

- Overfitting with too many predictors.
- Missing data or imbalanced classes in the `marital_status` variable.

3.1.9 Model justification

4 Results

Our results are summarized in **?@tbl-modelresults**.

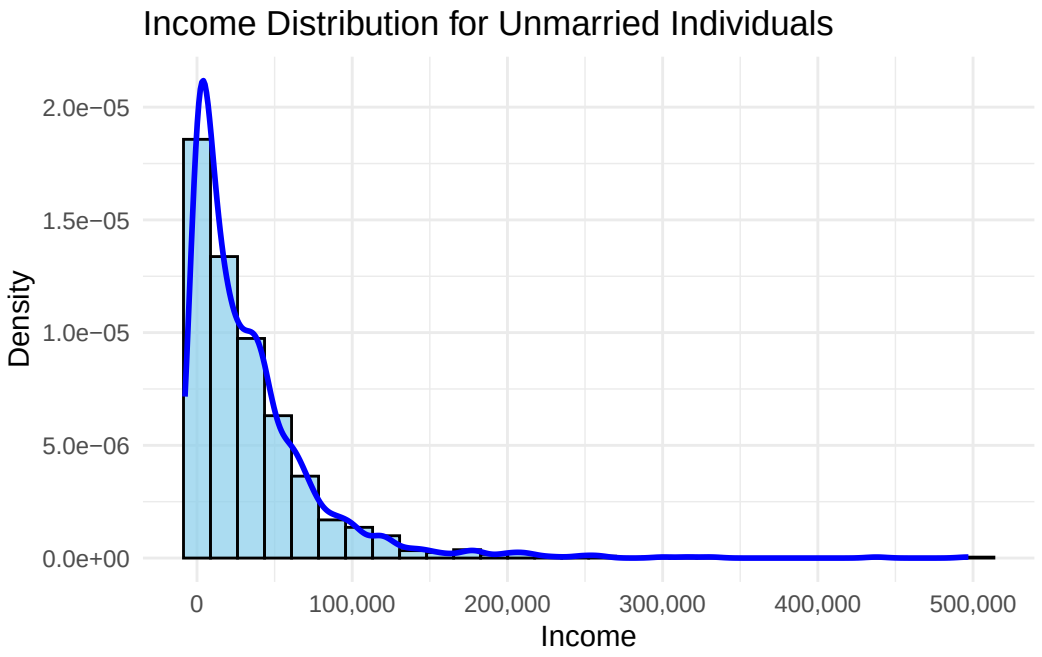


Figure 3: Income Distribution for Unmarried Individuals

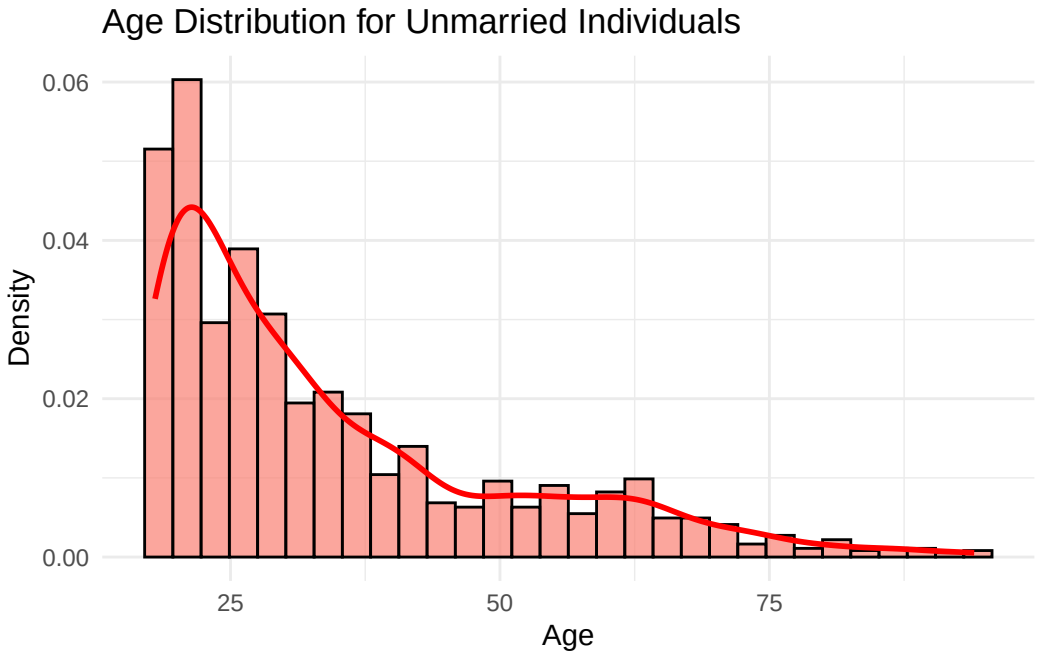


Figure 4: Age Distribution for Unmarried Individua

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

A.1 How to Extract 2022 ACS Data from IPUMS

To obtain data from IPUMS, we start by navigating to the IPUMS USA section and clicking on Get Data. Next, we go to the Select Sample section, where we uncheck the "Default sample from each year" option and instead select 2023 ACS. After selecting our sample, we proceed to add variables of interest. For individual-level data, we might add variables from the Person section. For example, under Demographic, we could include variables like AGE, and under Person, we could add SEX, RACE, INCTOT (total personal income) and EDUC (education attainment). Once our variables are selected, we click View Cart, then proceed by clicking Create Data Extract. At this point, we review our selections, change the Data Format to CSV, and submit our extract for processing. Then we saved it locally as usa_00001.csv.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

IPUMS. 2024. "IPUMS USA." https://usa.ipums.org/usa/#:~:text=IPUMS%20USA%20collects,%20preserves%20and%20harmonizes.

Müller, Kirill, and Jennifer Bryan. 2023. *here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Dewey Dunnington, and Apache Arrow Developers. 2023. *arrow: Integration to Apache Arrow*. https://arrow.apache.org/.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org/.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org/.

Wickham, Hadley, Dana Seidel, et al. 2023. *scales: Scale Functions for Visualization*. https://scales.r-lib.org/.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.