Analyzing Gender and Demographic Disparities: The Role of Education and Race*

Unpacking the Influence of Race on Educational Returns: Exploring Variations by Gender

Ricky Yuan

December 3, 2024

Gender disparities in demographic outcomes persist as a critical issue, reflecting the complex interplay of factors such as education, race, and social structures. This study employs Bayesian linear regression to examine how gender, educational attainment, race, and other variables interact to shape demographic distributions. The results reveal that disparities are not solely determined by gender but are also significantly influenced by levels of education and racial background. These findings emphasize the need for targeted interventions to address structural inequalities and improve our understanding of demographic patterns in socioeconomic contexts.

1 Introduction

Gender differences in demographic outcomes have been a significant problem in modern economies, reflecting inequalities in individual opportunities and the broader social structure. These disparities go beyond the gender pay gap and highlight the interplay between demographic inequality and social systems (Ponthieux and Meurs 2015). While much research has explored the gender pay gap, the underlying mechanisms—particularly how education quality, occupational choices, and social norms contribute to disparities in demographic variables such as age—remain insufficiently understood. Moreover, there is a lack of detailed quantitative analysis integrating these variables to account for intersectional factors like race and age. Solving this shortfall requires a robust framework to examine the interaction between demographic and socioeconomic variables. Bayesian learning models have shown useful strategy in this regard. They provide insights into the role of beliefs and demographic factors in shaping gendered disparities (Breen and Garcia-Penalosa 2002). This study aims to extend the existing literature by focusing on the factors that interact to shape demographic inequalities.

*Code and data are available at: https://github.com/RickyYuan666/Educationmain

The key to this study is a detailed examination of how gender, race, and education influence age distributions by integrating demographic and socioeconomic variables, including education level, race, and age. Further, this research uses Bayesian linear regression to quantify how these factors interact and contribute to demographic disparities. Existing research often relies on traditional statistical methods, which fail to capture parameter estimation uncertainty and the probabilistic nature of relationships. By adopting a Bayesian approach, this study addresses these methodological limitations and provides a more comprehensive understanding of the mechanisms driving gendered demographic inequalities. Also, this approach not only enables better parameter estimation but also offers a methodological basis for future economic research.

The results of our study show that age distributions vary by gender and race. Women tend to have a slightly higher average age compared to men, and racial differences further influence these distributions. However, the higher the level of education, the smaller the disparities in age distributions across genders. Significantly, these patterns are also influenced by the interaction between race and education levels, which shape demographic variations. These findings emphasize the impact of gender, education, and race on demographic outcomes, showing the persistent barriers that prevent equitable distributions across groups.

These findings not only enhance the issues that contribute to the persistence of demographic disparities but also emphasize the potential of education to solve these inequalities. As noted by Wrigley (2003), education plays a critical role in enhancing gender equality by enabling individuals to access better opportunities (Wrigley 2003). However, persistent inequalities in male-dominated fields suggest that more comprehensive measures are needed to eliminate structural biases. Therefore, this research identifies two key areas for future interventions: improving educational policies to address gender disparities in access and outcomes and reforming workplace structures to mitigate occupational isolation and bias.

The paper is structured to clearly analyze the demographic disparities and the role of education and race. Following an introduction by Section 1, Section 2 provides an overview of the dataset used in this study. It describes the core variables, such as gender, level of education, race, and demographic factors, as well as the rationale for their selection. Section 3 describes the Bayesian linear regression framework, emphasizing its advantages in addressing parameter uncertainty. Section 4 presents the results of the analyses, detailing how gender, education, and race interact to shape age distributions, with a particular emphasis on the mitigating role of education. Finally, Section 5 explores the broader implications of the findings, highlights the limitations of the current analysis, and suggests directions for future research.

1.1 Estimand

The main objective of this study is to estimate the impact of gender on age distributions, considering factors such as level of education, race, and transformed educational attainment. The

analysis quantifies how these variables interact to shape gendered differences in age distributions using Bayesian linear regression, focusing on the role of education and racial background in demographic disparities. By integrating probabilistic approaches to better capture uncertainty in parameter estimates, this study addresses gaps in existing research by providing a refined perspective on the factors driving gendered demographic inequalities.

2 Data

2.1 Data Description

The dataset utilized in this study was sourced from a publicly available repository on OSF (https://osf.io/48pqu/?view_only=), chosen for its comprehensive coverage of demographic and socioeconomic disparities and their associated factors. This dataset includes 1,000 observations, capturing critical variables such as gender, education level, race, age, and transformed educational attainment. Unlike other potential datasets, this repository offered the necessary depth for analyzing how demographic and educational factors contribute to gendered differences in age distributions. Alternative datasets were excluded due to their limited scope in categorizing educational attainment and racial backgrounds, which are crucial for the objectives of this research.

To ensure the dataset met the analytical goals, extensive pre-processing was conducted. Missing values were either imputed or removed, outliers were identified and addressed, and key variables were recoded for consistency and clarity. Gender was categorized into "Mulher" (female) and "Homem" (male), while educational attainment was ordered hierarchically, ranging from high school to postgraduate degrees. Ethnicity was recoded as "Branco" (white), "Preto" (black), "Pardo" (mixed), and "Indígena" (indigenous), allowing for a more nuanced analysis of intersectional impacts on age distributions. Age was retained as a continuous variable to account for differences in life stage and career trajectory.

After rigorous cleaning and preparation, this dataset serves as the foundation for testing how gendered differences in age distributions are shaped by demographic and socioeconomic factors. Also, the OSF repository ensures transparency and reproducibility, providing access to the dataset and associated processing codes, which align with the principles of open science. All data cleaning and analysis were performed in R (R Core Team 2023), leveraging its extensive statistical capabilities.

2.2 Data Source

Data for this research comes from the project "Money Illusion: A Replication of the 'Money Illusion' Effect in a Sample of Brazilian Volunteers", conducted in 2020 (Batistuzzo et al. 2020). This dataset provides information on demographic characteristics, educational attainment,

and racial distributions among Brazilian volunteers, as well as insights into psychological and behavioral patterns. The dataset, containing 1,000 observations and 31 variables, was accessed via the websie of OSF. It is favored over other available datasets due to its robust sampling methodology, comprehensive coverage, and relevance to the study of demographic and socioeconomic behaviors.

2.3 Measurement

The variables in this study were carefully chosen to reflect key phenomena influencing gendered demographic disparities in real-world contexts. Each variable was constructed or selected to capture essential aspects of demographic and educational dynamics relevant to the research question. For instance, age serves as the primary dependent variable, representing the demographic distribution across different groups. The dataset links age distributions to factors such as gender, education, and race, which are critical for understanding systemic inequalities.

The binary variable ("Mulher" for female and "Homem" for male) of gender captures the central focus of this study. It was designed to reflect societal gender distinctions that often result in demographic disparities. Educational attainment was ordered from high school to postgraduate degrees to emphasize how access to and quality of education influence demographic outcomes. Ethnicity categories ("Branco," "Preto," "Pardo," and "Indígena") reflect the intersectionality of race and gender, addressing how overlapping social categorizations influence age distributions. Finally, age, retained as a continuous variable, serves as a proxy for life stage and career trajectory, reflecting an individual's demographic profile.

These variables bridge the gap between real-world phenomena and analytical representation, enabling a rigorous examination of inequalities. To ensure the accurate representation of real-world dynamics, the dataset underwent pre-processing. For example, the hierarchical structuring of educational levels reflects its critical role in shaping demographic distributions, while ethnicity categories were refined to better represent Brazil's diverse population. These measurement decisions were designed to translate the complexity of societal patterns into a coherent analytical framework.

By integrating these thoughtfully constructed variables, this study establishes a strong foundation for connecting societal phenomena with empirical analysis, providing critical insights into the mechanisms of gendered demographic disparities.

- `tidyverse` (Wickham et al. 2021): For efficient data manipulation and visualization, streamlining the process of transforming and summarizing datasets.
- `ggplot2` (Wickham 2021): Used to create versatile and customizable visualizations tailored to the study's analytical needs.
- `dplyr` (Wickham, François, et al. 2021): Applied for intuitive data transformation, enabling effective handling of complex datasets.
- `bayesplot` (Gelman, Gabry, et al. 2021): For generating posterior predictive checks and diagnostic plots, supporting Bayesian model evaluation.

- `rstanarm` (Team 2021): Simplified the implementation of Bayesian models, providing an accessible framework for probabilistic regression analysis.
- `janitor` (Firke 2021): Facilitated data cleaning tasks, including renaming variables and standardizing dataset structures.
- `arrow` (Apache Arrow 2021): Enabled efficient reading and writing of large datasets, improving data accessibility and storage.
- `knitr` (Xie 2021): Used for generating dynamic reports, seamlessly integrating code, analysis outputs, and visualizations.
- ***Telling Stories with Data*** (Alexander 2023): Referenced for its insights into effectively presenting data and statistical analyses through clear narratives and visualizations.

Demographic-preview highlights the cleaned dataset's structure, showing variables such as gender, education level, race, and age. This dataset is integral to understanding how demographic and socioeconomic factors collectively shape gendered demographic disparities, as visualized and processed through R's advanced statistical tools.

2.4 Variables

Our analysis focuses on the following variables, with a specific emphasis on `idade` (age) as the dependent variable:

- `idade`: A continuous variable representing the age of individuals, serving as the dependent variable in our analysis. This variable provides insights into demographic distributions influenced by gender, education, and race.
- `gender`: The gender of the individual, coded as:
  - `Homem`: Representing men.
  - `Female`: Representing women.
    This variable is central to the study, analyzing its role in income disparities.
- `escolaridade_transf`: The transformed educational attainment variable, coded numerically to reflect hierarchical levels of education. This variable is critical in assessing how educational attainment influences demographic disparities.
- `race`: The racial background of the individual, coded as:
- `Branco` (White): Representing participants identifying as white.
- `Preto` (Black): Representing participants identifying as black.
- `Pardo` (Mixed): Representing participants of mixed race.
- `Indígena` (Indigenous): Representing indigenous participants. Including this variable allows for an intersectional analysis of race and gender in demographic patterns.

Detailed information about these variables and the data structure is presented in Demographic-preview, which outlines the first few records from the processed dataset.

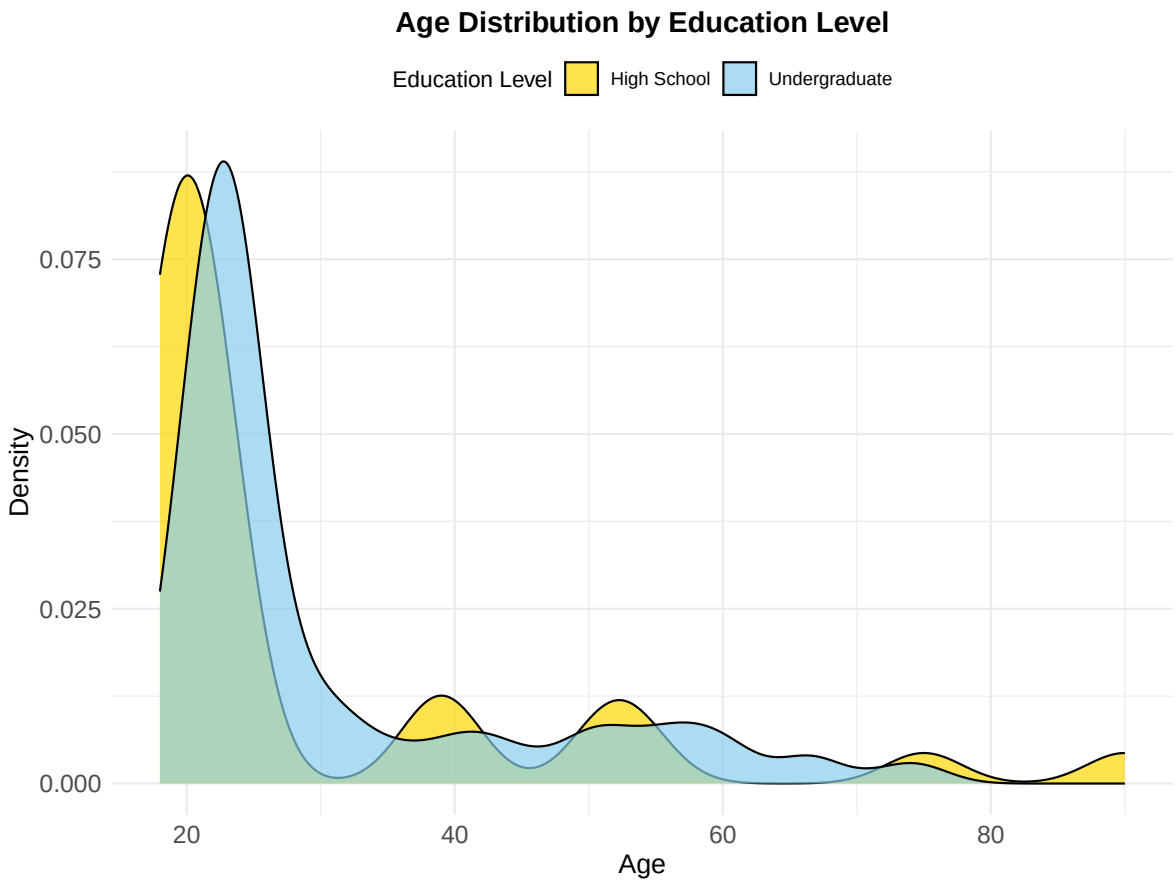

Figure 1: Density plot showing the age distribution across different education levels, highlighting variations in age groups.

Figure 1 illustrates the relationship between age distribution and education levels. The density plot highlights how age groups are distributed differently across education categories—High School, Undergraduate, and Postgraduate. The visualization reveals that individuals with High School education are concentrated in younger age groups, while those with Postgraduate education tend to cluster in older age ranges. Undergraduate participants show a broader distribution across the age spectrum. These findings suggest that educational attainment levels are often aligned with specific life stages, reflecting patterns of delayed or accelerated education pathways in different populations. This analysis provides insights into the demographic composition of each educational category and its implications for policy development and workforce planning.

Figure 2 illustrates the distribution of age across different education levels, highlighting the

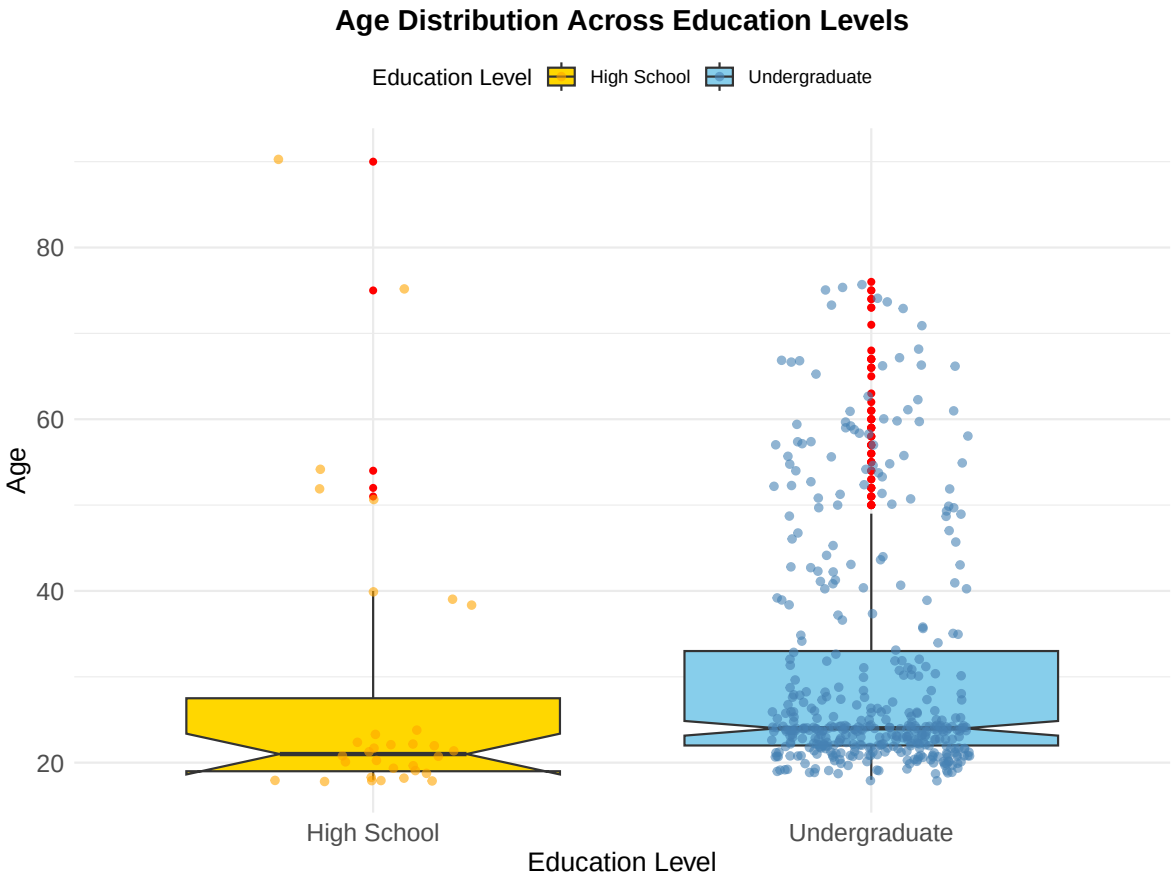

Figure 2: Boxplot with overlay showing the relationship between education levels and age, highlighting the age distribution across different educational attainments.

variation in age for individuals with high school and undergraduate degrees. The boxplot shows that the age distribution for undergraduate students is more varied, with a higher concentration of outliers at older ages. In contrast, individuals with only a high school education tend to cluster at younger ages, with fewer outliers. These findings suggest that higher education attainment may span a broader range of life stages, reflecting diverse career and life trajectories. This visualization emphasizes the role of education in shaping demographic distributions and provides insights into the intersection of age and education levels.

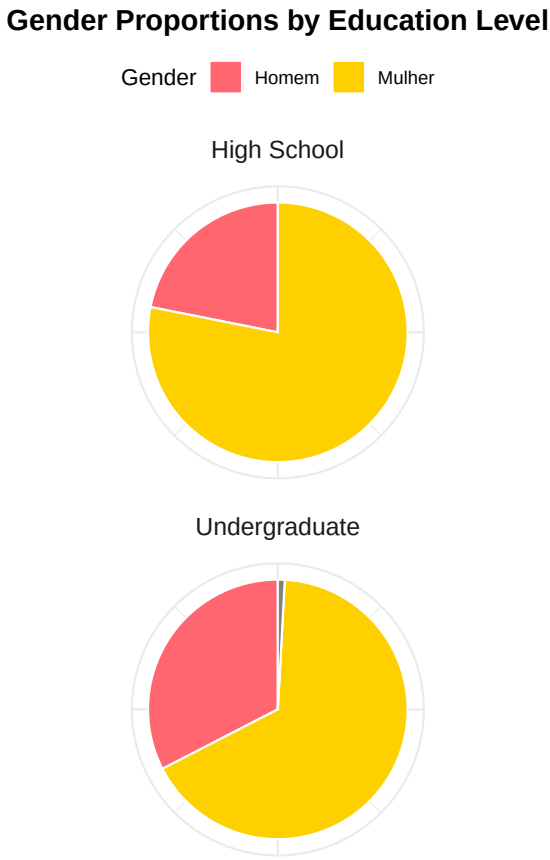

Figure 3: Pie chart showing the proportion of males and females within each education level.

Figure 3 presents the proportion of males ("Homem") and females ("Mulher") across different education levels using pie charts. Each chart represents a specific education level, with gender proportions illustrated by distinct colors. The visualization shows that at the high school level, females represent a larger proportion, whereas at the undergraduate level, the distribution between males and females is more balanced. The charts also highlight a small proportion of missing gender data in the undergraduate group. These findings underscore the relationship between gender and education levels, providing insights into demographic patterns in educational attainment.

2.5 Justification

The variables chosen for this study were carefully selected based on their relevance to understanding gender-based income disparities and the influence of education and career choices. Each variable provides critical insights into the socioeconomic factors driving income inequality.

- **Gender (`Gender`)**: Represents the primary demographic division for analyzing income disparities. This variable distinguishes between male and female participants to examine differences in earnings.
- **Education Level (`Education`)**: Captures the participants' highest level of education, categorized into high school, undergraduate, and postgraduate levels. This variable allows for exploring the role of educational attainment in moderating gender-based income disparities.
- **Career Field (`CareerField`)**: Reflects the occupational sector participants are involved in, categorized as humanities, STEM (science, technology, engineering, mathematics), and healthcare. This variable is crucial for understanding how occupational choices influence income levels by gender.
- **Income (`Income`)**: Serves as the dependent variable and measures the participants earnings. This variable provides a quantitative basis for evaluating disparities and testing the study's hypotheses.
- **Control Variables (`Age` and `Race`)**: Age is included as a continuous variable to account for variations in income over different career stages. Race is included as a categorical variable to control for potential intersectional effects on income disparities.

Variables such as **"Region"** or **"Family Structure"** were not included due to their lack of relevance to the primary research question or potential to introduce excessive complexity into the model. This focused approach ensures that the analysis remains interpretable and directly consitent with the study's goals.

By incorporating these variables, the study aims to provide a detailed analysis of how education and career field choices mediate income disparities between genders. The careful selection of these variables ensures that the research solves the central questions effectively while maintaining clarity.

3 Model

The variables chosen for this study were carefully selected based on their relevance to understanding gendered demographic disparities and the influence of education and race. Each

variable provides critical insights into the factors shaping demographic patterns, particularly age distributions. The model integrates the following key predictors:

- **Gender (`genero`)**: Represents the primary demographic division for analyzing disparities. This variable distinguishes between male ("Homem") and female ("Mulher") participants, allowing for comparisons in age distributions across genders.
- **Education Level (`escolaridade_transf`)**: Captures the participants' highest level of education, categorized into high school, undergraduate, and postgraduate levels. This variable enables the exploration of how educational attainment varies by gender and its role in shaping demographic profiles.
- **Age (`idade`)**: Serves as the dependent variable in this study, providing a continuous measure to analyze demographic distributions. Age reflects career stage and life trajectory, offering insights into variations across genders and educational levels.
- **Race (`decl_racial`)**: Reflects the participants' racial background, categorized into white ("Branco"), black ("Preto"), mixed-race ("Pardo"), and indigenous ("Indígena"). Including this variable allows for intersectional analysis of race and gender in shaping age distributions.

The Bayesian regression model is implemented using the brms package in R. This approach provides flexibility in defining priors, estimating posterior distributions, and conducting diagnostic checks. Priors for all parameters were informed by empirical research and domain knowledge, ensuring realistic constraints and interpretability. The model integrates weakly informative priors to regularize estimates while maintaining flexibility for the data to inform the results. Furthermore, the choice of a Bayesian framework allows the model to account for parameter uncertainty, providing a probabilistic interpretation of the predictors' effects. For instance, posterior distributions for gender coefficients offer insights into the likelihood and range of gender-based differences in age distributions.

3.1 Connection to Data Decisions

Including continuous age rather than grouped categories reflects the need to preserve data granularity and detect demographic patterns across different life stages. Similarly, the categorical classification of education levels and racial backgrounds ensures that the model captures key intersectional dynamics relevant to demographic disparities. These decisions align with theoretical frameworks on educational attainment and racial inequalities, ensuring that the model reflects real-world phenomena and provides meaningful insights into the factors influencing age distributions.

3.2 Model set-up

Let y_i represent the continuous variable denoting the **age** (`idade`) for the i-th observation. The predictors in the model include:

- β_1: The coefficient for the `escolaridade_transf` variable, which represents the highest level of education attained. This is an ordinal variable with categories:
  - *High School*
  - *Undergraduate*
  - *Postgraduate*
- β_2: The coefficient for the `gênero` variable, indicating the gender of the respondent. This is a binary variable with:
  - *Homem* (Male)
  - *Mulher* (Female)
- β_3: The coefficient for the `decl_racial` variable, representing racial self-identification. Categories include:
  - *Branco* (White)
  - *Preto* (Black)
  - *Pardo* (Mixed-race)
  - *Indígena* (Indigenous)
- β_4: The coefficient for the interaction between `escolaridade_transf` and `gênero`, capturing how the effect of education level on age varies by gender.

The linear predictor η_i for the i-th observation is defined as:

$$\eta_i = \beta_0 + \beta_1 \cdot \text{escolaridade\_transf}_i + \beta_2 \cdot \text{gênero}_i + \beta_3 \cdot \text{decl\_racial}_i + \beta_4 \cdot (\text{escolaridade\_transf}_i \times \text{gênero}_i) + \epsilon_i$$

where:

- β_0: The intercept, representing the baseline age for a reference group (e.g., male, high school, white).
- ϵ_i: Residual error, assumed to follow a normal distribution, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

3.2.1 Prior Distributions

Weakly informative priors are assigned to regularize the model:

$$\beta_j \sim \mathcal{N}(0, 10), \quad \sigma \sim \text{Cauchy}(0, 2)$$

This Bayesian regression model provides a probabilistic assessment of the relationships between demographic and educational predictors and age distributions. The interaction term allows the model to explore how gender modifies the effect of education on age, capturing nuanced dynamics in the data.

3.3 Prior distributions

In the Bayesian linear regression model implemented, weakly informative priors are applied to the model parameters to ensure robust and reliable inference. These priors provide a balance between regularization and flexibility, allowing the data to inform the results without imposing overly strong assumptions:

- **Intercept Priors**: A normal prior distribution is applied to the model's intercept with a mean of 0 and a standard deviation of 10. This choice stabilizes the baseline estimate for age without overly constraining its value.
- **Coefficient Priors**: The coefficients associated with the predictors are assigned normal prior distributions with a mean of 0 and a standard deviation of 2.5. This standard deviation is chosen to allow moderate effects while regularizing excessively large coefficients unless strongly supported by the data.

The priors are applied to the parameters associated with the following predictors: - `escolaridade_transf`: Capturing the ordinal levels of education (High School, Undergraduate, Postgraduate). - `gênero`: Representing the binary gender variable (Male, Female). - `decl_racial`: Reflecting the categorical racial groups (White, Black, Mixed-race, Indigenous).

- **Interaction Term**: Accounting for how the relationship between education and age differs by gender.

These priors ensure a balanced approach to capturing the effect of these predictors on age while avoiding overfitting to noise in the data.

The model was run in R using the brms package, a flexible framework for Bayesian regression modeling. The priors were carefully chosen to provide default regularization, ensuring the model remains interpretable while capturing meaningful patterns in the data. This approach allows for probabilistic insights into the relationships between demographic and educational predictors and age distributions.

4 Results

4.1 Model Justification

The coefficient estimates provide critical insights into the relationships between demographic and educational factors and age distributions. The gender coefficient (gênero) reveals notable differences, with females (Mulher) showing a positive mean coefficient. This indicates that females, on average, are slightly older in this dataset compared to males (Homem) when controlling for other factors. These results may reflect broader trends related to career trajectories, societal roles, or life circumstances that differ by gender.

The education level coefficient (escolaridade_transf) exhibits a positive association with age. Higher education levels, such as undergraduate and postgraduate degrees, are associated with older average ages. This finding aligns with expectations, as individuals pursuing advanced education tend to enter these programs later in life, particularly postgraduate studies. This trend emphasizes the importance of lifelong learning and its interaction with age-related career development.

The racial categories (decl_racial) demonstrate mixed associations with age. For example, individuals identifying as "Outro (Especificar)" display a significantly positive coefficient, suggesting they tend to be older on average compared to other racial categories. In contrast, coefficients for "Indígena" and "Branco" are not statistically significant, indicating minimal differences in age across these groups after controlling for education and gender. These results highlight the nuanced intersectionality of demographic variables and their impact on age.

The model's intercept represents the baseline average age for the reference group, which includes individuals who are male (Homem), identify as "Branco," and have the lowest educational level. This baseline allows for interpreting other coefficients relative to this group. For instance, the intercept value of approximately 23.7 suggests that the reference group has an average age of 23.7 years, providing a starting point for understanding how other predictors adjust this baseline.

These findings collectively illustrate how demographic and educational variables influence age distributions within the dataset. They underscore the importance of considering intersectional factors when analyzing age-related patterns in educational and career trajectories.

5 Discussion

This study applies Bayesian linear regression modeling to analyze the effects of gender, education level, and race on age distributions. The findings reveal that gender remains a key determinant, with females (Mulher) showing a slightly higher average age compared to males (Homem). This underscores the persistent structural differences in life trajectories, including varying educational timelines and societal roles. Education level demonstrates a strong positive

correlation with age, highlighting its role in shaping individual trajectories. Higher education levels, such as undergraduate and postgraduate degrees, are associated with older average ages, reflecting the extended timelines required for advanced studies. Furthermore, racial categories reveal nuanced patterns in age distributions. For instance, individuals identified as "Outro (Especificar)" are associated with significantly higher average ages, while other racial categories, such as "Branco" and "Indígena," show minimal differences. These results emphasize the interconnected factors shaping age-related demographic patterns and provide actionable insights for addressing disparities in education and career trajectories. By adopting an intersectional approach, this study contributes to a deeper understanding of how demographic and educational factors interact to influence age distributions.

5.1 Extensive Understanding of Target Selection

This study concentrates on variables that significantly influence age distributions, such as gender, education level, and race. These factors were selected for their relevance in reflecting real-world demographic and educational conditions. For example, higher education levels are consistently associated with older average ages, reflecting extended timelines for advanced studies, while gender and race expose differences in life trajectories that shape demographic patterns. By focusing on these dimensions, the research provides a comprehensive framework for understanding how structural and individual factors converge to shape age-related trends. Additionally, including these variables reflects an effort to align the analysis with broader societal trends, offering insights into the mechanisms influencing demographic and educational inequalities.

5.2 Strategic Implications of Variable Selection

Strategic Implications of Variable Selection: The deliberate inclusion of predictors like education level, gender, and race highlights the strategic intent to investigate age-related disparities at multiple levels. Education underscores the transformative impact of academic qualifications on life trajectories, with higher education levels often associated with older average ages due to extended timelines for advanced studies. On the other hand, gender and race expose deeply embedded societal patterns, illustrating how these identities intersect to shape demographic trends. The findings emphasize that addressing age-related inequalities requires multifaceted interventions, such as promoting equitable access to higher education, understanding gendered timelines in education and careers, and addressing intersectional challenges faced by underrepresented racial groups. Future research could explore additional variables, such as regional or cultural contexts, to build on these insights and broaden their applicability.

5.3 Weaknesses and Future Research Directions

Weaknesses and Future Research Directions: Despite its contributions, this study is subject to certain limitations. The dataset does not capture geographic variations or cultural contexts that could offer deeper insights into localized age-related trends. Moreover, excluding variables like family composition or socioeconomic status limits the scope of analysis. The relatively small sample size also constrains the ability to generalize findings across broader populations.

Future studies should address these gaps by incorporating contextual factors, such as regional education policies or the interplay between gender and education timelines across different demographic settings. Exploring these intersections could enrich our understanding of how age distributions vary across social and cultural contexts. Adopting advanced methodologies, such as hierarchical models or machine learning approaches, could also identify latent relationships and enhance predictive accuracy. These expansions would refine current findings and support the formulation of more targeted and practical solutions for addressing demographic and educational disparities.

5.4 Envisioning the Future of Historical Military Analysis

Envisioning the Future of Historical Military Analysis: Future research could incorporate variables such as geographic location and cultural contexts to explore how regional demographics and social norms impact age distributions. Additionally, analyzing the intersection of demographic factors, like the combined influence of gender and education levels, could uncover more nuanced patterns in life trajectories. Expanding the dataset or adopting hierarchical Bayesian models or advanced statistical techniques may also reveal hidden relationships, offering deeper insights into the structural drivers of age-related disparities.

5.5 The Value of Strategic Insights

The Value of Strategic Insights: This study emphasizes the importance of understanding the key factors behind age-related disparities, like education, gender, and race. Education stands out as a driver of life-course timing, showing how higher qualifications are consistently associated with older average ages, reflecting extended educational and professional trajectories. At the same time, the persistent effects of gender and race reveal structural patterns that policies must address to ensure equitable access to education and opportunities. These findings provide actionable insights for tackling demographic inequities and guiding strategies to support diverse educational and career pathways.

Appendix

A Model details

A.1 Data cleaning notes

We began by loading the raw dataset using the read_csv function from the readr package. To ensure the data was suitable for analysis, we removed the first row, which contained metadata rather than actual data values. This step was followed by renaming the column names using the make.names function from base R, ensuring consistent and readable column names.

Next, we addressed missing data by calculating the proportion of missing values in each column. Columns with more than 80% missing data were excluded from the dataset to maintain data quality. We also transformed categorical columns into factors using the mutate and across functions, ensuring a consistent data format for non-numeric variables. Duplicate rows were identified and removed to eliminate redundant information.

To enhance usability, numeric columns such as escolaridade_transf (education level) and idade (age) were checked for outliers and normalized where necessary. Additionally, we ensured that categorical variables, such as gênero (gender) and decl_racial (race), had clearly defined levels to avoid inconsistencies in subsequent analysis.

Finally, to organize the cleaned dataset effectively, we implemented a check-and-create process to verify the existence of an output folder structure. The final cleaned dataset was saved in both Parquet and CSV formats to provide compatibility with various tools and workflows.

These steps streamlined the raw data into a clean, analysis-ready format, minimizing potential errors and enhancing the dataset's usability for subsequent analysis.

A.2 Survey Methodology

The data collected for this study were used to analyze the factors affecting age distributions, focusing on variables such as gender, level of education, and race. Each survey targeted a different sample to capture a wide range of demographic and socioeconomic characteristics, ensuring the representation of critical populations relevant to the research questions.

To obtain a balanced dataset, a stratified sampling technique was used. Respondents were categorized into groups based on demographic factors such as age, gender, and education level. This method ensures proportional representation and minimizes potential bias due to underrepresenting specific groups, such as older individuals pursuing postgraduate degrees or racial minorities with advanced education levels. Further, stratification also reflects population demographics, enhancing the generalizability of the results. Post-survey adjustments were made to correct differences between the sample and the broader population.

Given the problems associated with nonresponse to the survey, we used a follow-up method to improve response rates. We incentivized respondents to participate in the survey and took steps to minimize the bias introduced by nonresponse. For example, demographic information about nonrespondents was analyzed to ensure that no significant biases could affect the results.

The questionnaire was designed to be targeted to maximize the relevance of the responses to the study objectives. The questions were concise and structured to provide clear insights into the relationship between age distributions and the selected predictors. This design ensured that the data collected were representative and directly applicable to the Bayesian linear regression model used in the analysis.

Overall, the data collection and processing methods employed provide a foundation for exploring the structural drivers of age-related disparities and offer actionable insights into the interplay of demographic and socioeconomic factors.

A.3 Idealized Survey

Section 1: Demographics What is your gender?
a) Male
b) Female
c) Non-binary
d) Prefer not to say

What is your age?
a) Under 18
b) 18–24
c) 25–34
d) 35–44
e) 45–54
f) 55–64
g) 65 or older

What is your race/ethnicity?
a) White
b) Black
c) Asian
d) Indigenous
e) Mixed
f) Other (please specify): _____
g) Prefer not to say

Section 2: Education Background What is your highest level of education?
a) Primary school
b) Secondary school
c) Undergraduate degree
d) Postgraduate degree
e) Professional/Doctorate degree

What is your area of study or specialization?
a) Humanities
b) STEM (Science, Technology, Engineering, Mathematics)
c) Health sciences
d) Business or Economics
e) Arts or Design
f) Other (please specify): _____

Section 3: Career and Employment What is your current employment status?
a) Employed full-time
b) Employed part-time
c) Self-employed
d) Unemployed
e) Student
f) Retired

What is your current monthly income (in BRL)?
a) Less than 1,000 BRL
b) 1,000–2,499 BRL
c) 2,500–4,999 BRL
d) 5,000–7,499 BRL
e) 7,500–9,999 BRL
f) 10,000 BRL or more
g) Prefer not to say

What type of employment contract do you currently have?
a) Permanent contract b) Temporary contract c) Freelance/No formal contract d) Not applicable

---

Section 4: Housing and Living Conditions Do you own or rent your residence?
a) Own

b) Rent
c) Live with family without paying rent
d) Other (please specify): _____

How many people contribute to the household income?
a) Only me
b) 2 people
c) 3 people
d) More than 3 people

Section 5: Opinions and Perceptions How satisfied are you with your current financial situation?
a) Very satisfied
b) Somewhat satisfied
c) Neutral
d) Somewhat dissatisfied
e) Very dissatisfied

How attractive do you find your current job in terms of career growth?
a) Very attractive
b) Somewhat attractive
c) Neutral
d) Not very attractive
e) Not at all attractive

Do you feel that your race or ethnicity has impacted your career opportunities?
a) Yes, positively
b) Yes, negatively
c) No impact
d) Prefer not to say

Do you believe your education level has significantly influenced your career path?
a) Yes
b) No
c) Not sure

References

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

Apache Arrow. 2021. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Breen, Richard, and Cecilia Garcia-Penalosa. 2002. "Bayesian Learning and Gender Segregation." *Journal of Labor Economics* 20 (4): 899–922.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gelman, Gabriel, Jonah Gabry, et al. 2021. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot.

Ponthieux, Sophie, and Dominique Meurs. 2015. "Gender Inequality." In *Handbook of Income Distribution*, 2:981–1146. Elsevier.

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

Team, Stan Development. 2021. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Wickham, Hadley. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2021. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://tidyverse.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wrigley, Julia. 2003. *Education and Gender Equality*. Routledge.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.