My title*

My subtitle if needed

First author Another author

November 24, 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 Overview

Our data (**shelter?**)….

We use the statistical programming language R (R Core Team 2023)….Our data (**shelter?**)….

The IPUMS USA database (Ruggles et al. 2024) is one of the largest collections of microdata from population censuses globally. Supported by organizations such as the National Institutes of Health and the University of Minnesota, it includes data from the American Community Survey (ACS) and other census programs. This resource provides detailed individual-level data, allowing researchers to address specific social, economic, and demographic questions. A

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

key feature of IPUMS USA is its capacity to create customized datasets by selecting variables based on research needs, minimizing the inclusion of irrelevant data.

For this study, data from the 2023 ACS sample in IPUMS USA were used. The selected variables include STATEFIP (state code), SEX (gender), AGE (age), RACE (race), EDUC (educational attainment), EMPSTAT (employment status), UHRSWORK (usual hours worked per week), and INCWAGE (wage and salary income). These variables cover geographic location, demographic characteristics, education, and employment conditions, providing a broad basis for examining factors influencing wage income.

Other databases, such as IPUMS CPS (Flood et al. 2024) and the National Longitudinal Surveys (Bureau of Labor Statistics 2023), were considered but deemed less suitable for this analysis. IPUMS CPS provides detailed labor market data but has a smaller sample size and limited geographic detail. The National Longitudinal Surveys track specific population groups over time but lack broad representation, making them less ideal for cross-sectional studies. In contrast, the 2023 ACS data in IPUMS USA offers larger population coverage and a diverse range of variables, making it a better choice for analyzing wage determinants.

2.2 Measurement

This study utilizes sample data from the 2023 ACS. The dataset is based on a 1% random national sample, ensuring broad representativeness. It includes individuals living in private households as well as those in group quarters, such as student dormitories and care facilities. Weights were applied to the data to account for potential biases introduced by the sampling design and nonresponse.

Data collection for the ACS involved structured questionnaires covering a wide range of topics, such as demographics, education, employment, and income. For example, UHRSWORK (usual hours worked per week) captures labor input based on respondents' reports of their average weekly work hours. The smallest geographic unit in the dataset is the Public Use Microdata Area (PUMA), which contains at least 100,000 residents and falls within state boundaries. This structure protects respondent privacy while supporting regional analysis.

While the ACS methodology ensures data reliability and extensive coverage, variables like income and work hours rely on self-reported responses, which may introduce minor inaccuracies. Nonetheless, the use of sampling weights and robust design make the dataset a reliable foundation for studying wage income and its associated factors. A detailed explanation of the survey methodology and sampling design is provided in the appendix.

2.3 Dataset Description

Table 1: Table 1: Categories of SEX Fully Displayed

education_level	UHRSWORK	INCWAGE	region	age	gender	race_group
Above_Bachelor	40	105000	Northeast	49	Female	Asian
Bachelor	40	73000	South	29	Male	White
Below_High_School	45	30000	West	45	Female	Other
High_School	40	37000	West	27	Male	Other
Some_College	10	1200	South	22	Male	Black

The cleaned dataset includes the following variables. UHRSWORK and INCWAGE are retained from the raw data, while the other variables are newly constructed. During the data cleaning process, only individuals aged 18-65 and those with an employment status of "Employed" were included. This ensures the dataset focuses on studying the wage determinants of the employed labor force. Below is a description of each variable and how it was constructed:

UHRSWORK: Reports the number of hours per week the respondent usually worked if they were employed during the reference period.

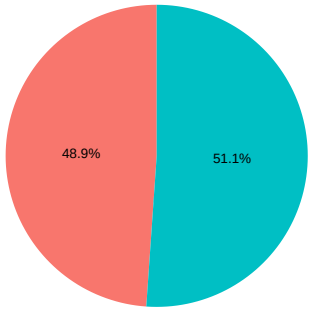
INCWAGE: Represents each respondent's total pre-tax wage and salary income earned as an employee during the previous year.

region: Constructed from the STATEFIP variable to represent geographic regions. Valid state codes (1–56) were grouped into four traditional U.S. regions: Northeast, Midwest, South, and West. Each state was assigned to its corresponding region for further regional analysis.

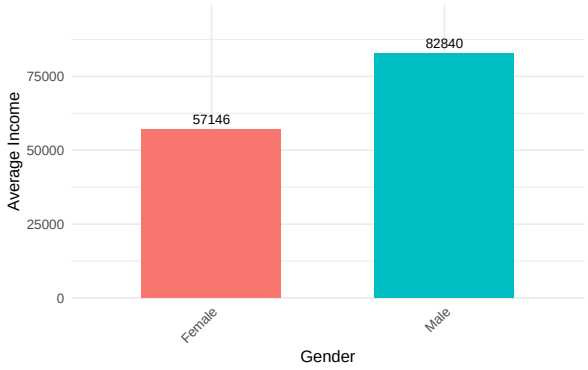
education_level: Derived from the EDUC variable, which indicates respondents' educational attainment as measured by the highest year of school or degree completed. It is grouped into five categories: Below High School (includes all levels below high school), High School (completed grade 12), Some College (completed 1-2 years of college), Bachelor's Degree (completed 4 years of college), and Above Bachelor (more than 5 years of higher education).age: Filtered from the AGE variable to include only respondents aged 18 to 65. This variable records the respondent's age in years as of their last birthday.

gender: Based on the SEX variable, with original codes recoded as "Male" and "Female."

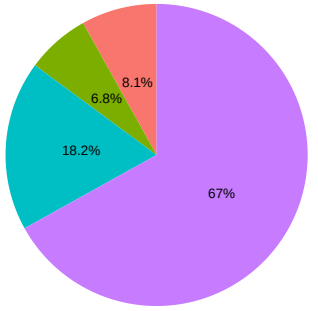
race_group: Constructed from the RACE variable and grouped into four categories: White, Black, Asian, and Other. White and Black remain consistent with the original classifications. Asian includes Chinese, Japanese, and Other Asian categories, while Other includes all other races and mixed-race groups.



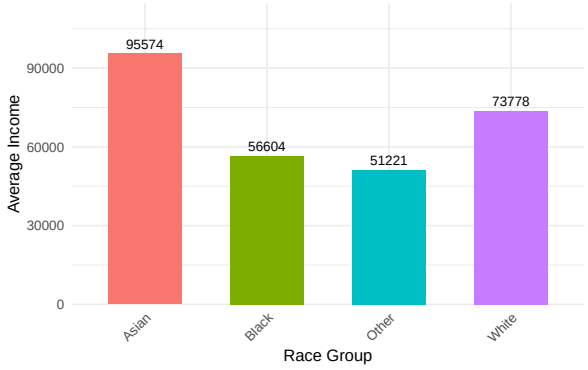
(a) Gender Proportion



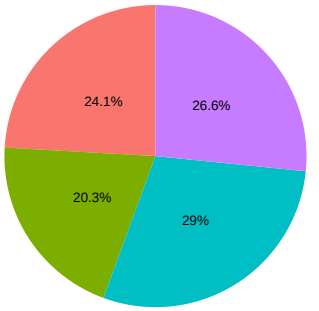
(b) Average Income by Gender



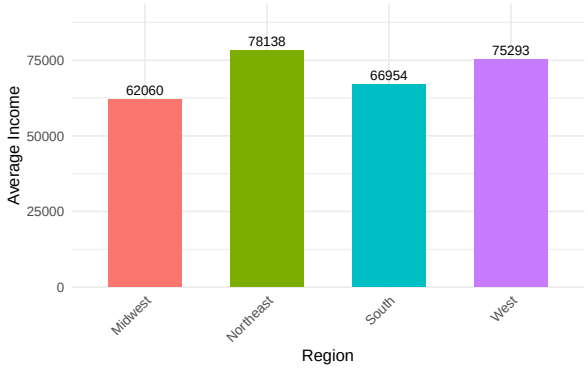
(c) Race Proportion



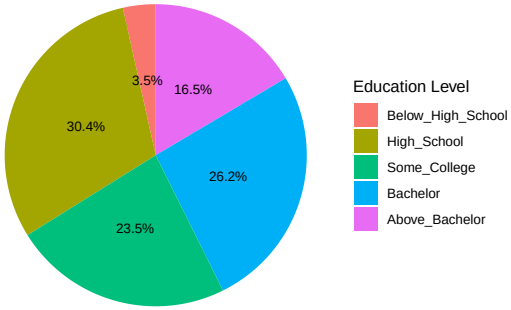
(d) Average Income by Race



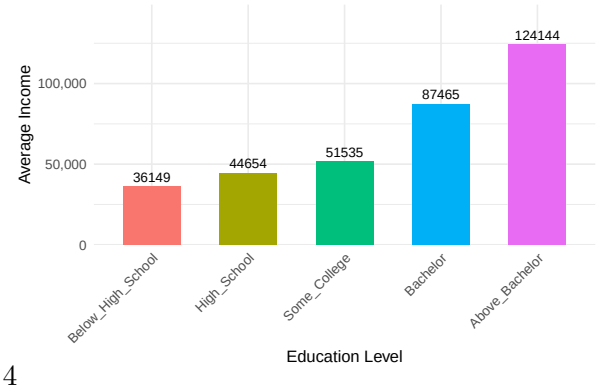
(e) Region Proportion



(f) Average Income by Region



(g) Education Proportion



(h) Average Income by Education Level

Figure 1: Demographics Distribution and Average Income by Group

2.4 Data Results

Figure 1 illustrates the distribution of demographic characteristics and their relationship with average income across gender, race, region, and education level. Pie charts display the proportional distribution of each group, while bar charts show corresponding average income levels, providing a straightforward depiction of these relationships.

Chart (a) shows the gender distribution: females account for 48.9% of the population, and males 51.1%, indicating a nearly equal split. Chart (b) shows that males earn an average income of $82,840, compared to $57,146 for females.

Chart (c) presents the racial composition, with White individuals forming the largest group at 67%, followed by Black individuals at 18.2%, Asians at 6.8%, and other racial groups at 8.1%. Chart (d) shows that Asians have the highest average income at $96,574, followed by White individuals at $73,778. Black individuals and other racial groups earn $51,221 and $56,604, respectively.

Chart (e) shows the regional distribution. The West represents the largest share at 29%, followed by the Northeast (26.6%), the South (24.1%), and the Midwest (20.3%). Chart (f) indicates that the Midwest has the lowest average income at $62,046, while the Northeast has the highest at $78,138. The South and West have average incomes of $66,954 and $75,293, respectively.

Chart (g) shows education levels: 16.5% of the population has less than a high school education, 30.4% has completed high school, 23.5% has some college education, 26.2% holds a bachelor's degree, and 9.5% has education beyond a bachelor's degree. Chart (h) highlights the connection between education and income. Those with less than a high school education earn $36,149 on average, while individuals with a bachelor's degree or higher earn $87,465 and $124,144, respectively. Income increases steadily with higher levels of education.

Figure 2 examines the relationship between income, work hours, and age, incorporating both individual-level data and aggregated trends.

Chart (a) is a scatter plot of work hours versus income, with each point representing an individual. The red quadratic regression line indicates a positive correlation between income and work hours, where longer hours generally lead to higher income. However, the regression line primarily reflects a linear trend, suggesting a stable relationship between income and work hours without significant signs of deceleration. Additionally, income variability increases with longer work hours, particularly beyond 50 hours per week, where fluctuations become more pronounced. Chart (b) illustrates the relationship between work hours and average income. While income generally increases with work hours, especially between 20 and 50 hours per week, it becomes unstable beyond 50 hours.

Chart (c) is a scatter plot of age versus income, with the red quadratic regression line showing that income increases with age before eventually declining. Chart (d) depicts the trend of average income by age. The line chart shows that average income steadily rises with age,

(a) Work Hours vs Income

(b) Average Income by Age

(c) Age vs Income

(d) Average Income by Work Hours

Figure 2: Income Trends by Age and Work Hours

peaking in the late 40s and early 50s, and gradually declines as individuals approach retirement age.

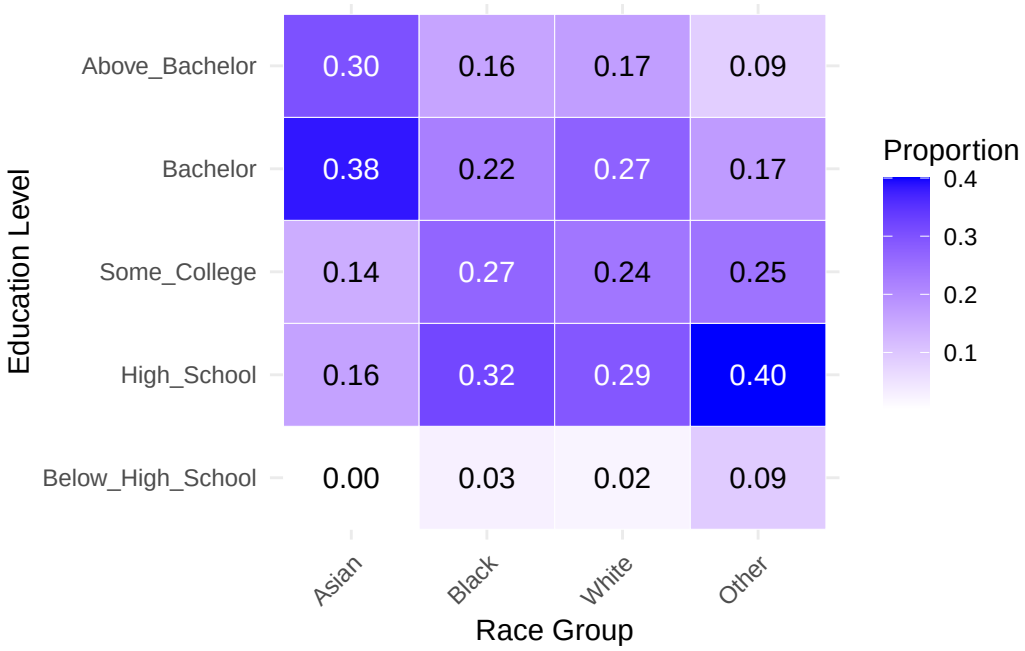


Figure 3: Proportion of Education Levels by Race Group

The chart illustrates the proportion of different education levels across racial groups. The horizontal axis represents racial groups, including "Asian," "Black," "White," and "Other," while the vertical axis represents education levels, ranging from "Below High School" to "Above Bachelor." The intensity of blue indicates the proportion, with darker shades representing higher values.

Among Asians, the proportion of highly educated individuals is relatively high, with 38% holding a bachelor's degree and 30% having education beyond a bachelor's degree, while the proportion of those with "Below High School" education is 0%. In the Black group, "High School" accounts for the largest proportion at 32%, followed by "Some College" at 27%, while those with "Above Bachelor" degrees make up only 16%. In the White group, "High School" accounts for 29%, with 27% holding a bachelor's degree and 24% having "Some College" education, while 17% have "Above Bachelor" degrees. Among the "Other" group, 40% have a "High School" education, 25% have "Some College," 17% hold a bachelor's degree, and only 9% have "Above Bachelor" degrees.

Overall, the distribution of education levels varies across racial groups. Asians have a higher proportion of highly educated individuals compared to other groups, while "High School" education is more prevalent among Black and "Other" groups. This reflects the complex relationship between race and educational background.

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

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

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.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

References

Bureau of Labor Statistics, U. S. Department of Labor. 2023. "National Longitudinal Survey of Youth 1979 Cohort, 1979-2020 (Rounds 1-29)." Columbus, OH: Produced; distributed by the Center for Human Resource Research (CHRR), The Ohio State University.

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, et al. 2024. "IPUMS CPS: Version 12.0 [Dataset]." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D030.V12.0.

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

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0 [dataset]." Minneapolis, MN: IPUMS, University of Minnesota. https://doi.org/10.18128/D010.V15.0.