Factors influencing Total Income in the United States*

Analyzing Factors influencing Total Income in the United States of America from 2019 to 2022

Wen Han Zhao

December 1, 2024

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

Table of contents

*Code and data are available at: https://github.com/younazhao/Income-Influencing-Factors.git.

1 Introduction

Income inequality is a central concern in economic research, reflecting disparities in access to opportunities, resources, and outcomes. In the United States, total income is influenced by a variety of factors, including personal, professional, and demographic characteristics. This paper explores the relationship between total income and predictors such as marital status, sex, educational level, labor force participation, and occupation. By employing statistical modeling techniques, we aim to uncover the extent to which these variables shape income levels and contribute to broader income disparities.

The estimand of this study is the expected total income for individuals based on their social, personal and occupational characteristics. Specifically, the analysis seeks to quantify how factors such as marital status, sex, educational attainment, participation in the labor force, and occupational category predict variations in total income across the U.S. population. This includes identifying both independent and among these variables.

Preliminary results indicate that educational attainment is one of the strongest predictors of total income, with higher levels of education yielding significantly greater earnings. Marital status also plays a notable role, with married individuals generally earning more than their single counterparts. Gender disparities in income persist, with men earning more on average than women across most occupational categories. Labor force participation and occupation further highlight critical differences, as full-time workers and those in higher-skilled professions consistently report higher incomes.

Understanding the drivers of income disparities is essential for designing policies that promote economic equity and mobility. By identifying how marital status, gender, education, labor force participation, and occupation contribute to income levels, this research offers actionable insights for policymakers, employers, and educators. Addressing these disparities not only fosters greater fairness but also enhances the overall economic productivity and social cohesion of the United States.

To do this, total income was analyzed in Section 2 Section 2. The prediction from the model has show accuracy in lower total income group. The model has shown difficulties in predicting higher total income. Section 3 Section 3 has further explanation on the Bayesian Generalized

Linear Model. Section 4 Section 4 show the coefficient beta from the model. Section 5 Section 5 focus on the weakness of this study and further possible improvements. Lastly, Section 6 **?@sec-appendix** contains methodology to perform a survey to collect data for Total Income.

2 Data

2.1 Overview

The Survey of Consumer Finances (SCF), conducted in 2022 covering the period 1989 to 2022 by the Board of Governors of the Federal Reserve System of the United States (Board of Governors of the Federal Reserve Board 2023), provides a comprehensive snapshot of household financial conditions across the nation. This dataset includes detailed information related to total income, capturing key predictors such as sex, education level, occupation category, salary income, and income from other sources. By offering rich, granular data, the SCF serves as an invaluable resource for analyzing the factors influencing income distribution and economic disparities within the United States.

2.2 Measurement

The survey contained a panel element over 2 periods. Respondents to the 1983 survey were re-interviewed in 1986 and 1989. Respondents to the 2007 survey were re-interviewed in 2009. In order to ensure the representativeness of the study, respondents are selected randomly in order to attempt families from all economic strata.

2.3 Outcome variables

The histogram represents the distribution of our outcome variables total income among survey participants. From the graph, we can see that the graph shows slightly right skewed but mostly symmetric by following a bell shape curve. It shows a mean centered 11 to 12, indicating that most individuals in the survey have incomes within this range. The spread of the graph shows approximately from 8 to 16 of the log income and contains only a few extreme outliers at the tails. Overall, the distribution of log income, the outcome variable, follows a normal distribution.

Quantile-quantile plot is included to compare the log-transformed income data' distribution to a theoretical normal distribution. The data should expect to follow closely to the blue dashed line to indicate a normal distribution. From the graph, we can see that the middle of the plot lies closely to the line which indicate that the data follows a normal distribution. The right tail

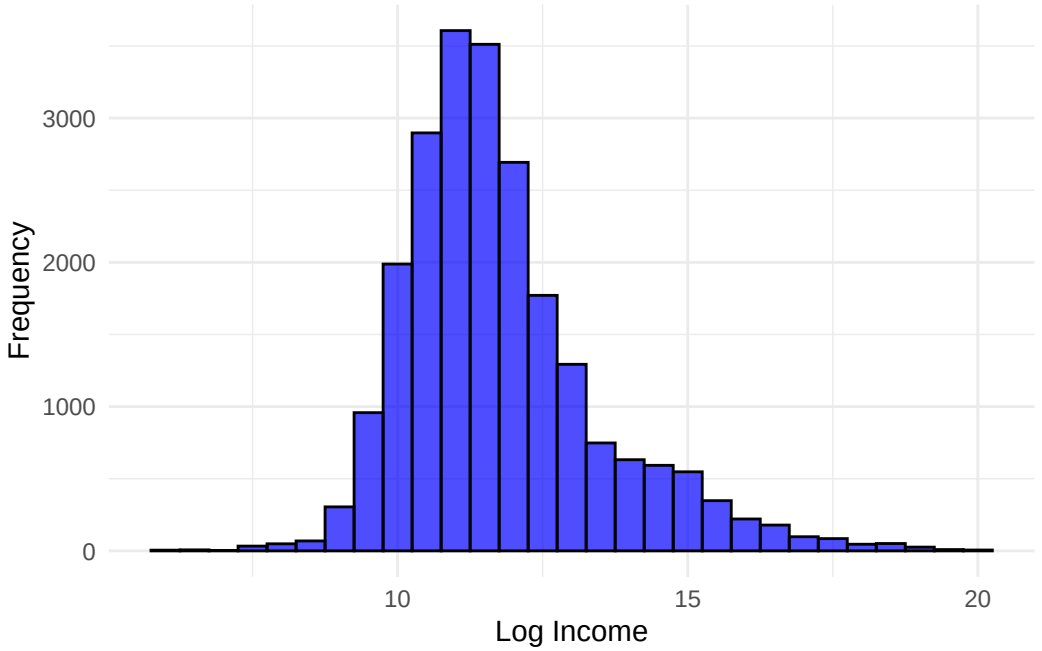

Figure 1: Distribution of Total Income

have various points deviate above the line, which suggest a heavier right tail. The deviation form the right tail is expected, since higher income are more random and harder to predict.

2.4 Predictor variables

Predictors variables selected are Age, Sex, Marital Status, Education Level, Labour Force, Work Status, and Occupation category.

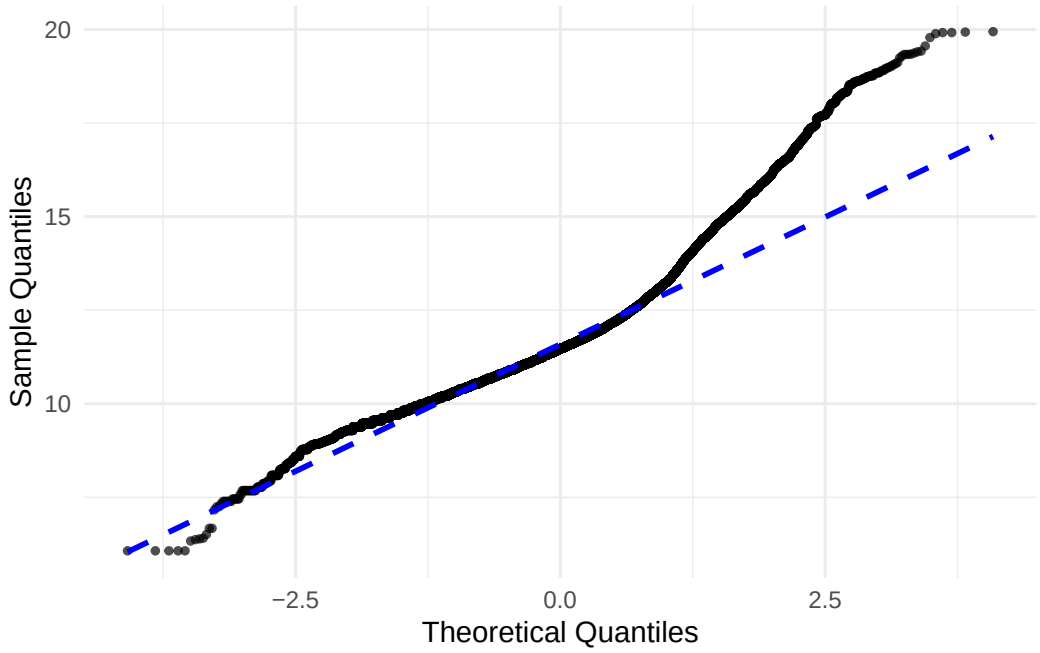

Figure 2: Comparison of Income Distribution using QQ Plot

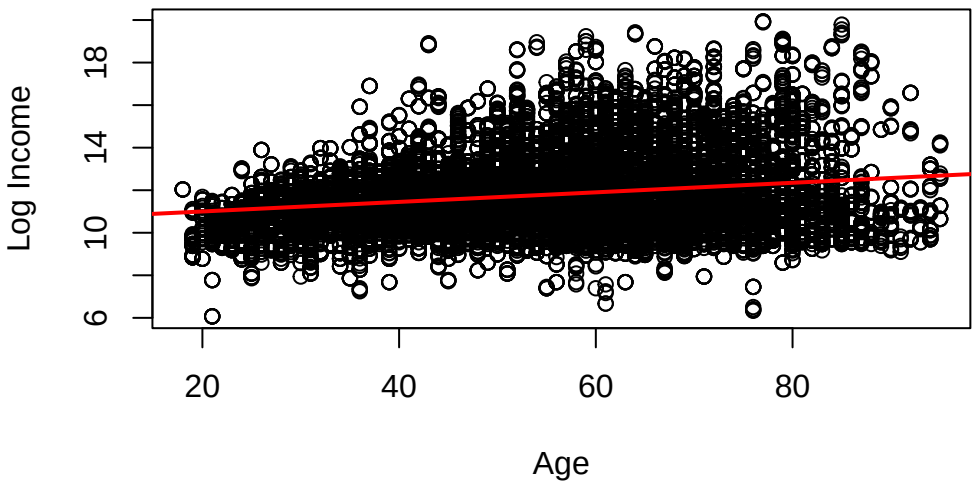

Figure 3: Log Income by Age

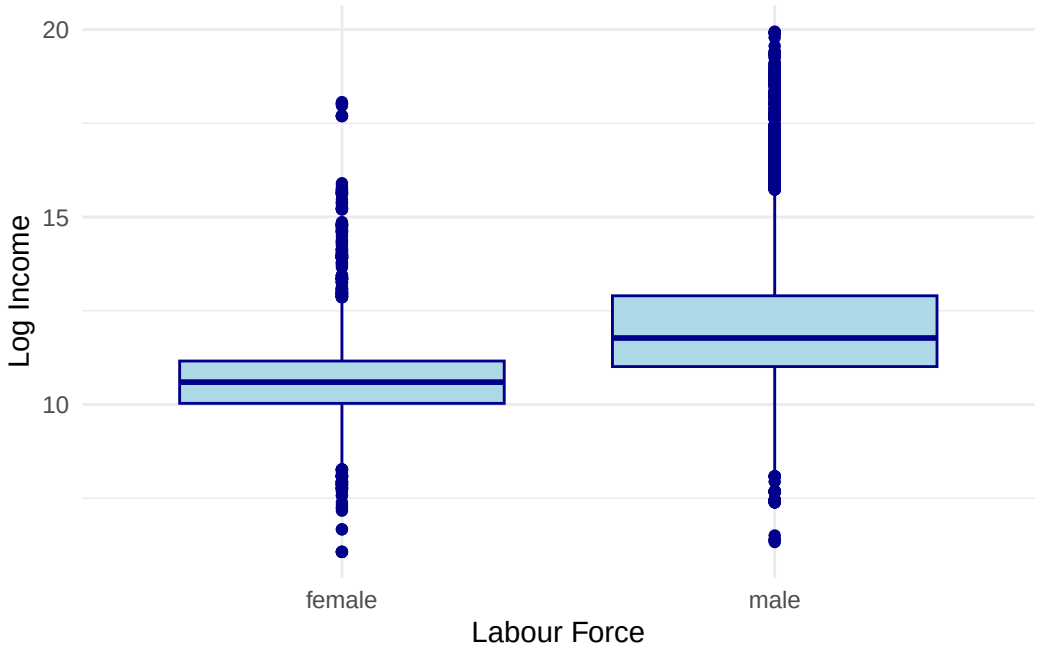

Figure 4: Log Income by Sex

Figure 5: Log Income by Marital Status

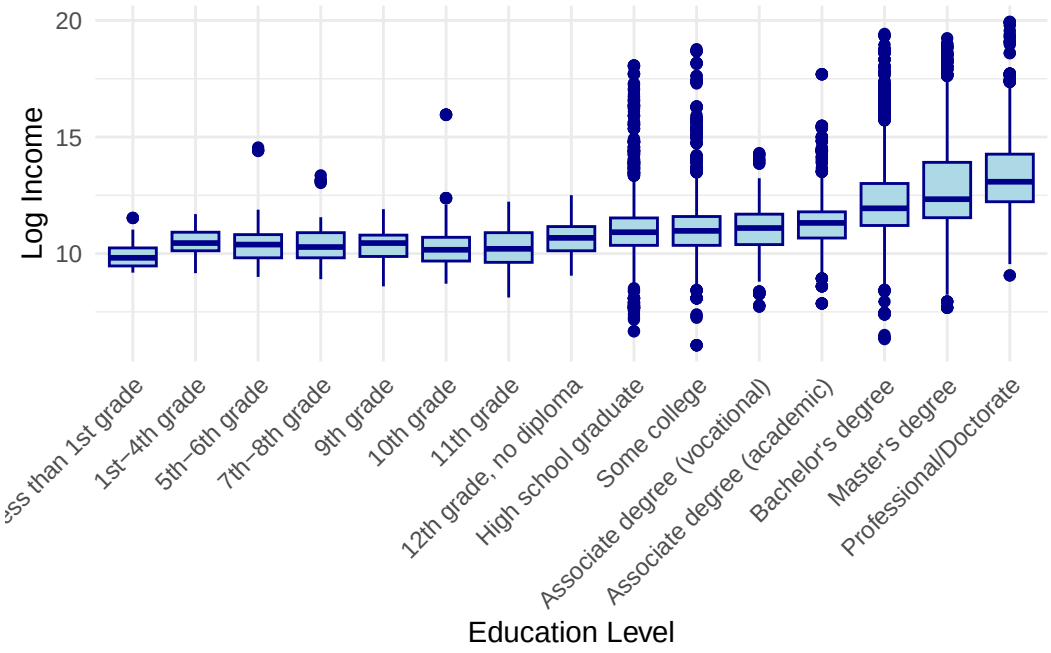

Figure 6: Log Income by Educational Level

Figure 7: Log Income by Labour Force

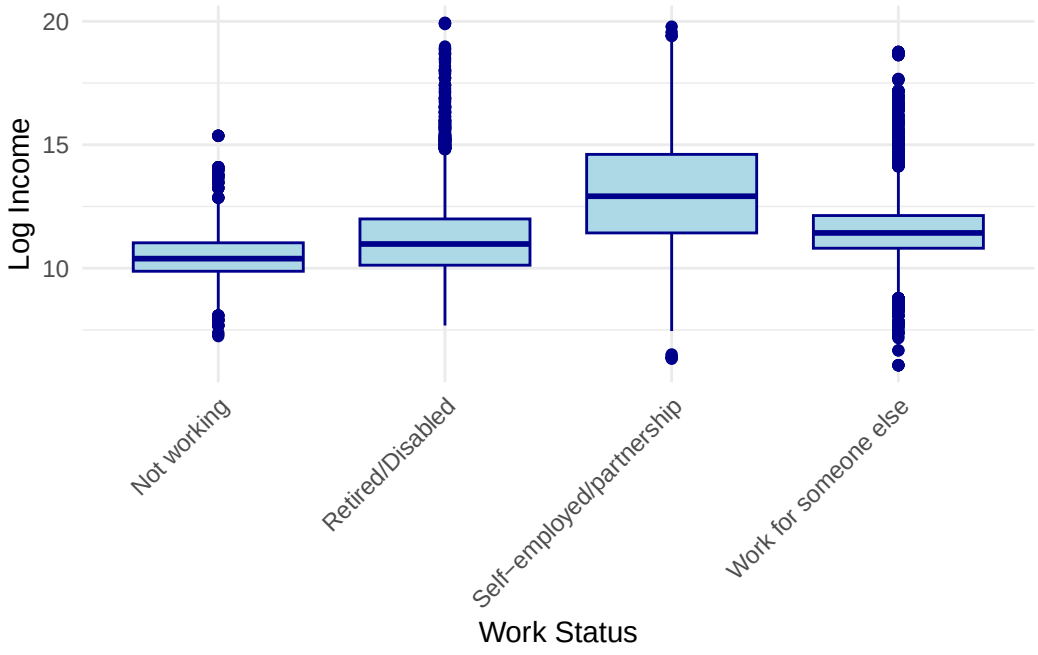

Figure 8: Log Income by Work Status

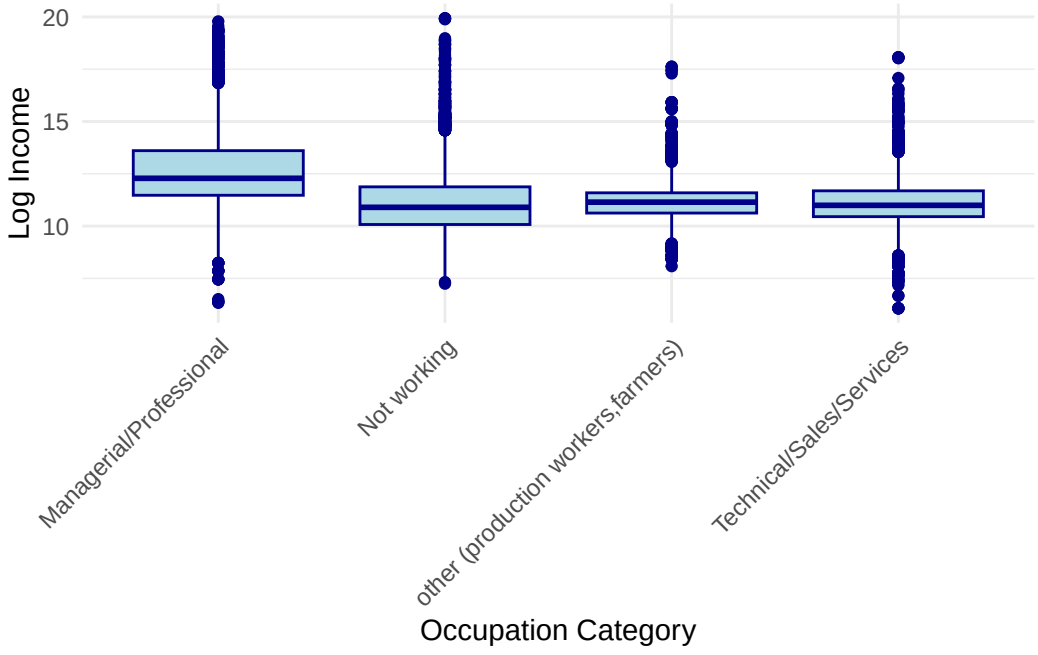

Figure 9: Log Income by Occupation Category

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 Total Income for each individual participated in the survey in the United States at 2022. The predictors are categorical variables with different levels explained in Section 2-.

$$
\begin{aligned}
\log(y_i) &= \beta_0 + \beta_1 \cdot \text{Sex} + \beta_2 \cdot \text{Education Level} + \beta_3 \cdot \text{Marital Status} \\
&\quad + \beta_4 \cdot \text{Labor Force} + \beta_5 \cdot \text{Work Status Category} + \beta_6 \cdot \text{Occupation Classification} + \epsilon, \\
\beta_0 &\sim \text{Normal}(10, 5), \\
\beta_1, \beta_2, \ldots, \beta_6 &\sim \text{Normal}(0, 2.5), \\
\epsilon &\sim \text{Normal}(0, \delta^2), \\
\delta &\sim \text{Exponential}(1).
\end{aligned}
$$

3.2 Linear Model

We have started with the linear model for predicting Total Income based on the selected variables. The model set up is:

$$
\begin{aligned}
\hat{\text{income}} &= \beta_0 + \beta_1 \cdot \text{Sexmale} + \beta_2 \cdot \text{Education Level} + \beta_3 \cdot \text{Marital Status} + \beta_4 \cdot \text{Labor Force} \\
&\quad + \beta_5 \cdot \text{num\_work} + \beta_6 \cdot \text{num\_occu}
\end{aligned}
$$

Table 1

Variable	Estimate	Std. Error	t-value	Pr($\gt \mid t \mid $)	Significance
(Intercept)	-348940	870243	-0.401	0.6884	
Sexmale	668026	263287	2.537	0.0112	*
educ	227466	33274	6.836	8.35e-12	***
married	-1103395	234718	-4.701	2.6e-06	***
lf	116525	353617	0.33	0.7418	
num_work	1892834	144769	13.075	$<$ 2e-16	***

num_occu	-1273936	123257	-10.336	< 2e-16	***
Residual Standard Error	1236000				
Multiple R-squared	0.02223				
Adjusted R-squared	0.02198				
F-statistic	86.3				
p-value	< 2e-16				

Table for Linear Model Results

3.3 Model justification

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`. The prior intercept have been set to $N(10, 5)$ because the histogram Figure 1 indicates a mean near 10-12 with variation of 5. In addition, various predictors would have strong correlation with our outcome variable Total Income. By setting the prior intercept would capture well the additional knowledge compared to the default priors from `rstanarm` package.

4 Results

Our results are summarized in Table **??**.

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.

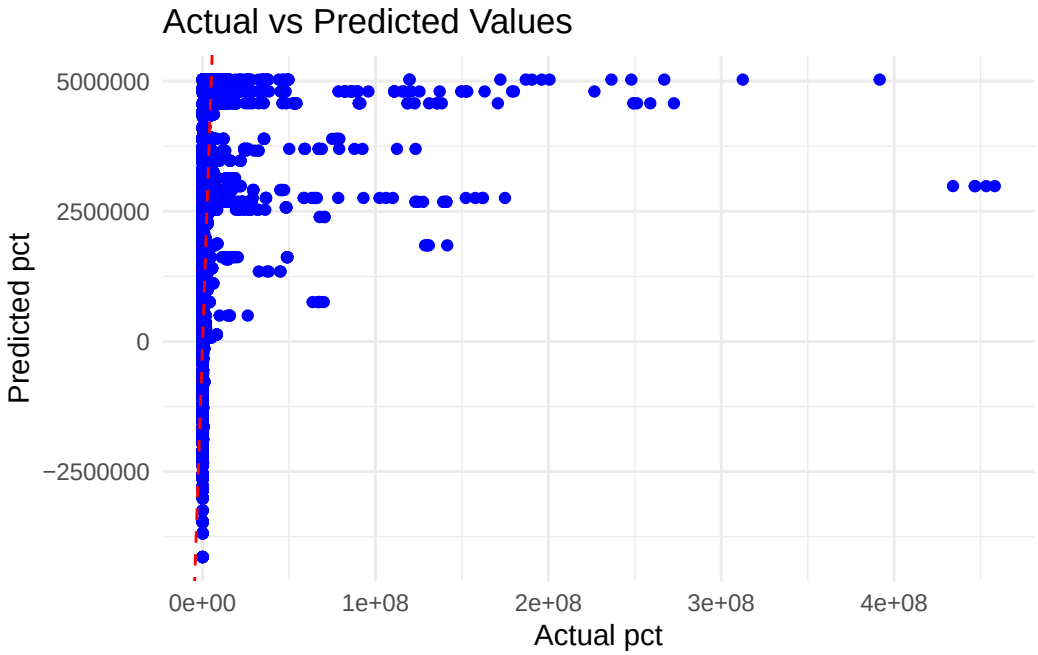

Figure 10: Linear Model has limited predictive accuracy for Total Income

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…

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

References

Board of Governors of the Federal Reserve Board. 2023. "2022 Survey of Consumer Finances." Board of Governors of the Federal Reserve System. https://doi.org/10.17016/8799.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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