

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

This paper develops a Bayes generalized linear model to predict the Total Income of United States of America through a survey conducted from 2019 to 2022. By analyzing factors such as Age, Sex, Marital Status, Education Level, Labour Force, Work Status and Occupation Category. By integrating these factors, we are able to see several trends and differences in each of those variables in order to best predict the Total Income in each individuals. Our results show that Male workers with higher education level at a managerial position will have higher Income. This model offers a comprehensive tool for understanding the different categories in each individual group and improving the accuracy of Total Income predictions.

Table of contents

1	Introduction	2
2	Data	3
2.1	Overview	3
2.2	Measurement	3
2.3	Outcome variables	4
2.4	Predictor variables	6
3	Model	13
3.1	Model set-up	13
3.2	Linear Model	13
3.3	Log Transformation of Linear Model	15

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

3.4	Bayes Model	16
3.5	Bayes Model on Log Transformation	17
3.6	Model justification	18
4	Results	18
5	Discussion	21
5.1	Summary of Findings	21
5.2	Second discussion point	21
5.3	Third discussion point	21
5.4	Weaknesses and next steps	21
A	Appendix	22
B	Model details	22
B.1	Posterior predictive check	22
B.2	Diagnostics	22
	References	24

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.

The data gathering and analysis is done in R (R Core Team 2023) with the following packages: Wickham et al. (2019), Xie (2022), Developers (2022), Wickham et al. (2022a), Wickham et al. (2022b), Wu (2022), and Goodrich et al. (2022).

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.

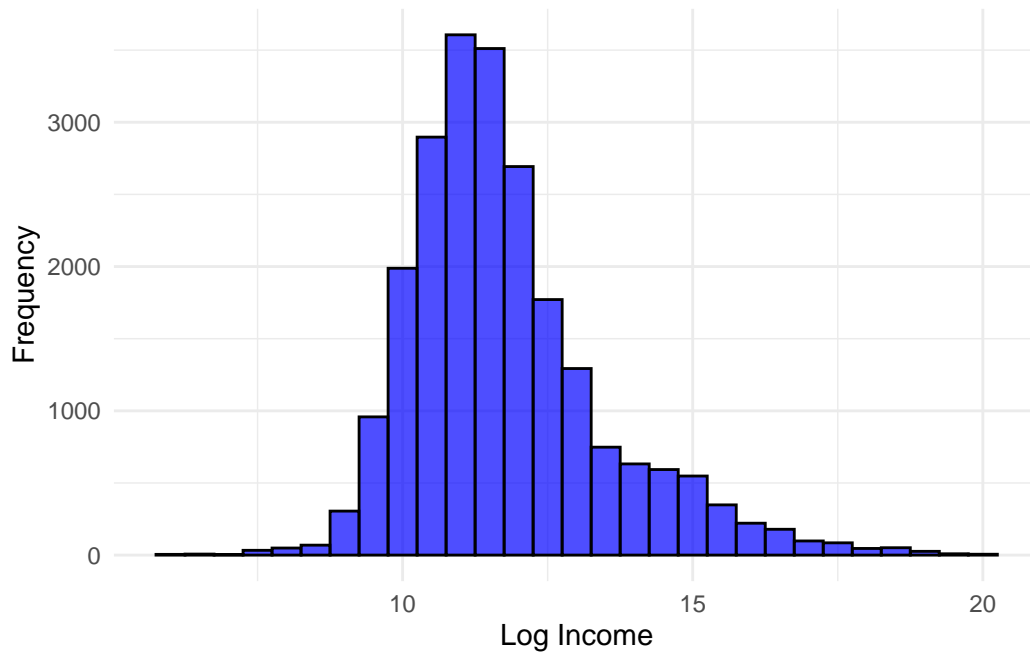


Figure 1: Distribution of Total Income

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

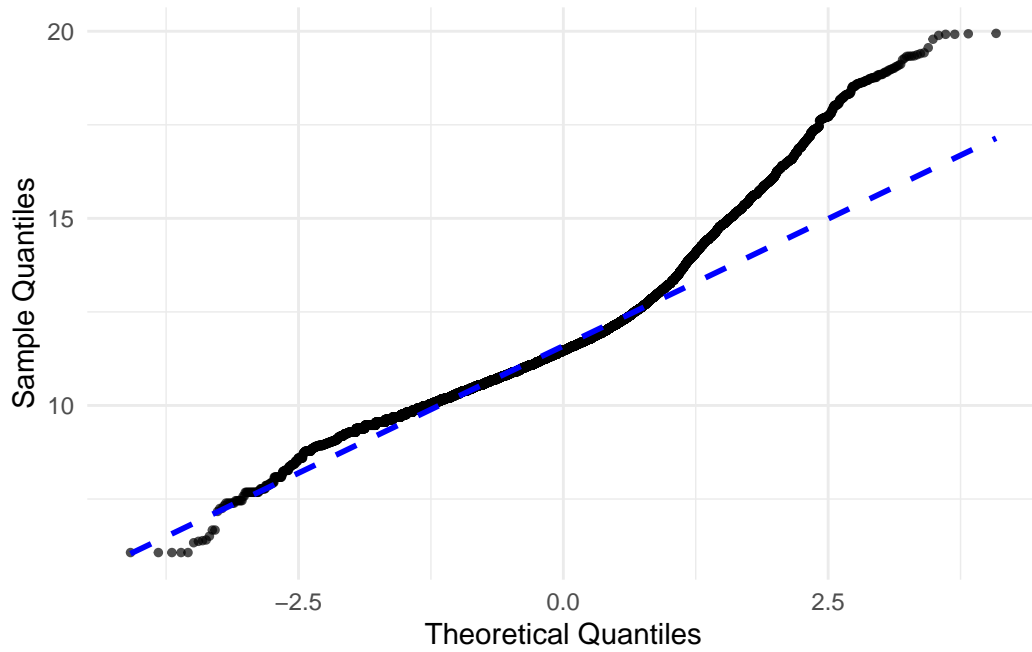


Figure 2: Comparison of Income Distribution using QQ Plot

2.4 Predictor variables

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

In Figure 3, the graph shows a scatter plot with Age on the x-axis and Log Income on the y-axis, with individual data points represented as black circles. A red regression line is overlaid, indicating the linear relationship between age and log-transformed income. The plot suggests a positive correlation between age and income, where income increases as age increases. However, the relationship appears relatively weak, as the data points are widely spread around the regression line, indicating substantial variability in income across different ages. Despite this, the general upward trend is clear, showing that older individuals tend to have higher log-transformed income. Therefore, we decided to not use this variable in our rating.

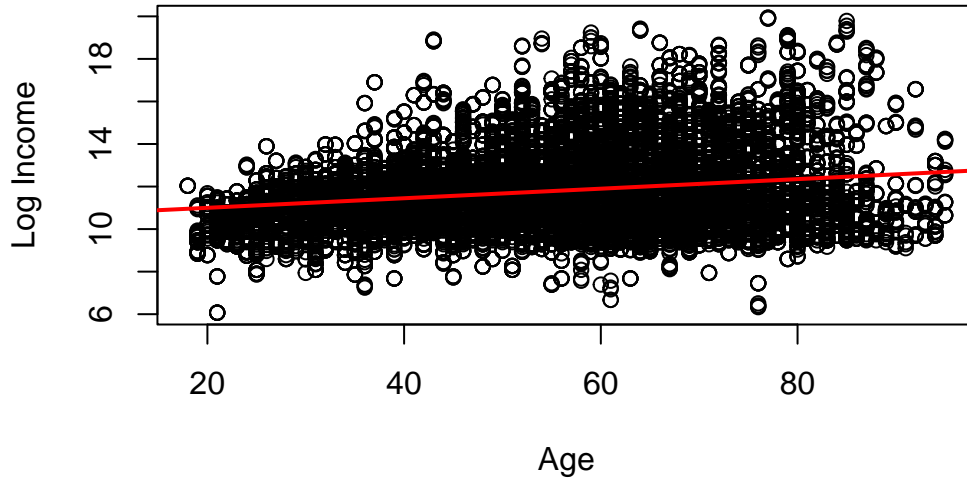


Figure 3: Log Income by Age

In Figure 4, the graph is a boxplot comparing log-transformed income (Log Income) between females and males in the labour force. The y-axis represents the log of income, while the x-axis distinguishes between females and males. The boxplots show that both groups have a similar range of income, with the male group exhibiting a wider spread and higher income values, as indicated by the larger box and more significant presence of outliers. The female group's boxplot is more concentrated around the lower range of log income, with fewer outliers, suggesting that females generally earn less in comparison to males in this dataset. The median for females is lower, and the distribution is less variable than for males.

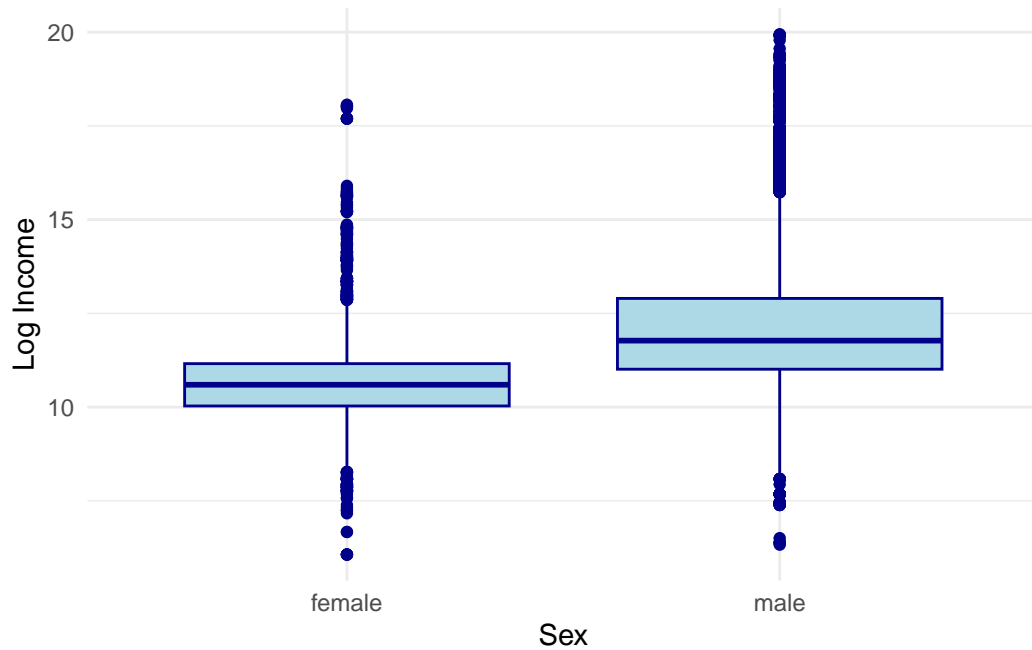


Figure 4: Log Income by Sex

In Figure 5, the boxplot compares log-transformed income (Log Income) between married and not married individuals in the labour force. The y-axis represents the log of income, while the x-axis distinguishes between the two marital status groups. The boxplot for married individuals shows a higher median and wider distribution of income, with more outliers present, indicating a broader range of income levels and higher earnings overall. In contrast, the not married group has a lower median income, and the distribution is more tightly clustered around lower values, suggesting that individuals who are not married tend to have lower log income compared to their married counterparts.

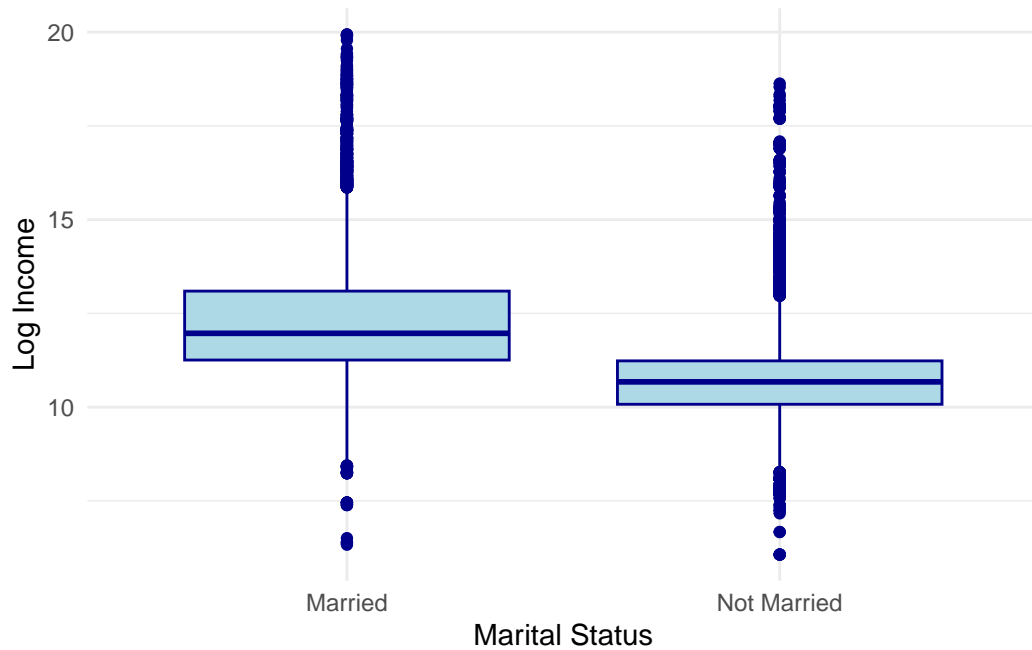


Figure 5: Log Income by Marital Status

In Figure 6, the boxplot compares log-transformed income (Log Income) across different education levels. The y-axis represents log income, while the x-axis shows various levels of education, ranging from less than 1st grade to professional/doctoral degrees. The plot reveals a clear positive trend: as education level increases, the median log income tends to rise, with individuals holding a Master's degree or Professional/Doctorate showing the highest median incomes. The spread of income also increases with higher education levels, particularly for individuals with Bachelor's degrees and beyond, where there are more outliers. This indicates that individuals with higher education levels generally have greater income, but the income distribution becomes more varied at the higher education levels.

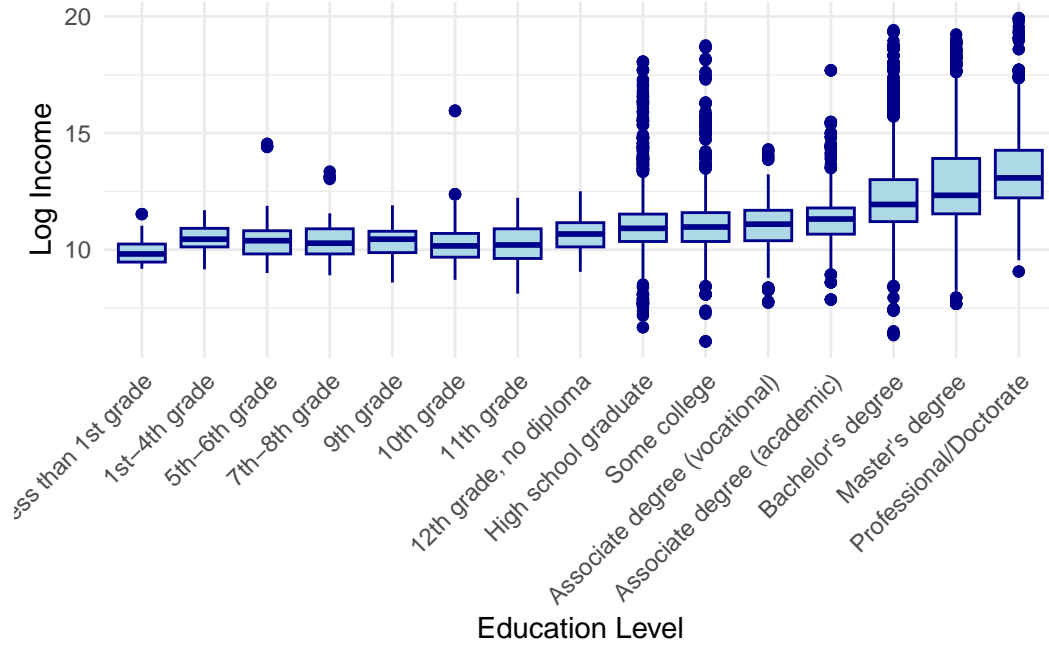


Figure 6: Log Income by Educational Level

In Figure 7, the boxplot compares log-transformed income (Log Income) between individuals who are working and those who are not working at all in the labour force. The y-axis represents log income, while the x-axis differentiates between the two groups. The plot shows that individuals working in some way generally have higher log incomes, with a wider spread and more outliers, particularly at the upper end. In contrast, those not working at all tend to have lower log incomes, with a narrower distribution and fewer outliers. The median for the “working in some way” group is higher, indicating that employment is associated with higher income. The presence of more outliers in the working group suggests greater income variability among those employed.

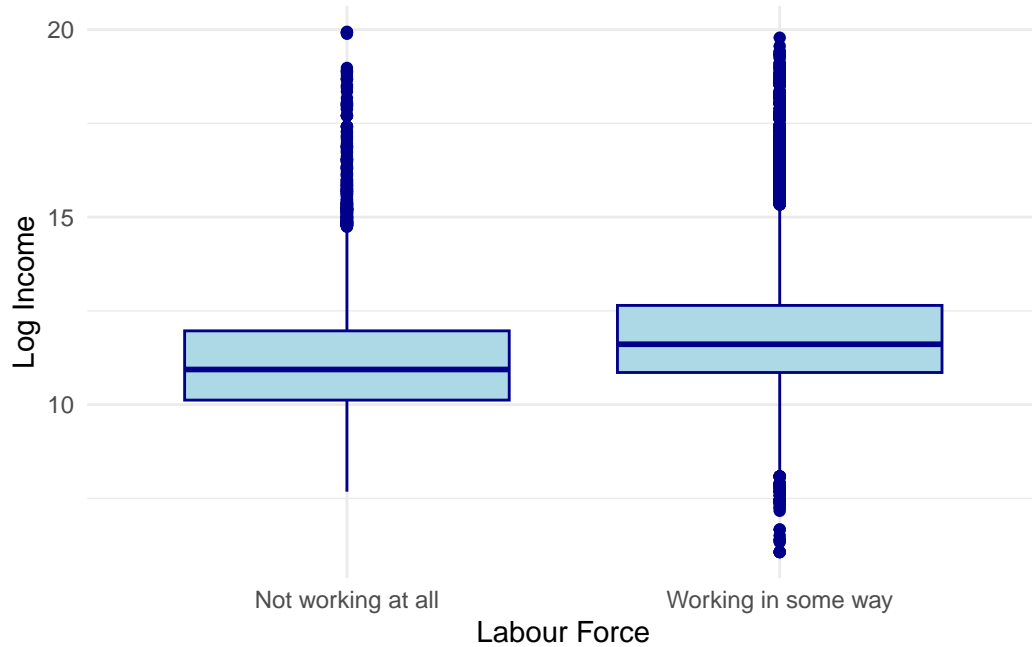


Figure 7: Log Income by Labour Force

In Figure 8, the boxplot compares log-transformed income (Log Income) across four different work statuses: Not working, Retired/Disabled, Self-employed/Partnership, and Work for someone else. The y-axis represents log income, while the x-axis differentiates between the work status categories. Individuals who work for someone else tend to have the highest median log income, with a wider spread and more outliers, indicating greater income variability. Self-employed/partnership individuals also show higher median incomes but with a slightly smaller spread. Those who are retired/disabled have a narrower distribution, with a lower median income, and the not working group has the lowest median income, with a similar narrow distribution. The plot clearly demonstrates that employment, particularly in self-employment or working for someone else, is associated with higher log income compared to being retired, disabled, or not working.

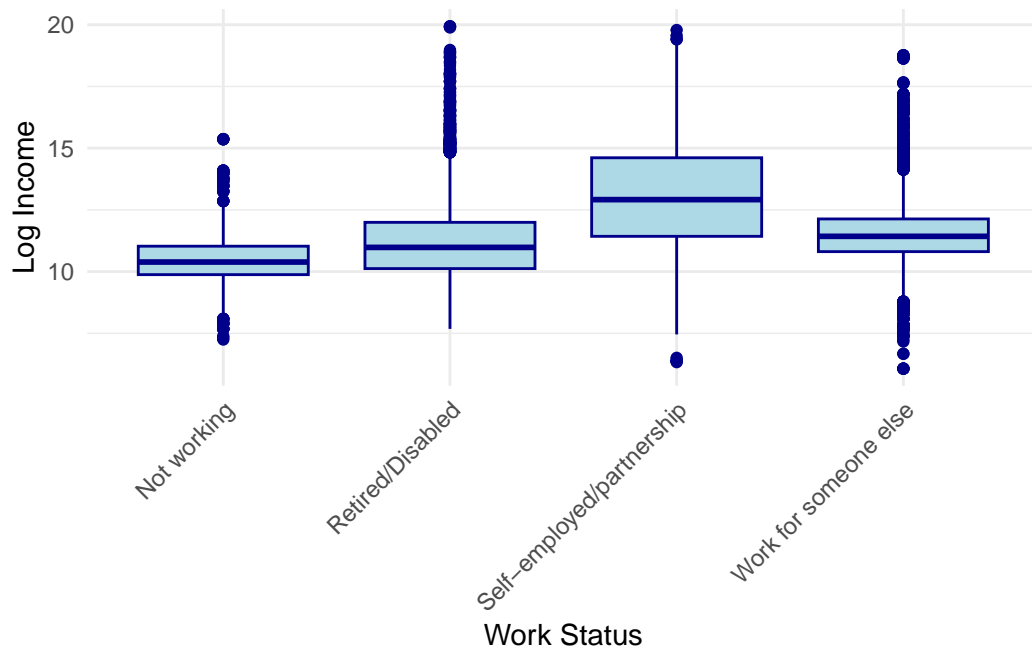


Figure 8: Log Income by Work Status

In Figure 9, the boxplot compares log-transformed income (Log Income) across four occupation categories: Managerial/Professional, Not working, Other (production workers/farmers), and Technical/Sales/Services. The y-axis represents log income, and the x-axis differentiates between the occupation groups. Individuals in the Managerial/Professional category have the highest median income, with a wider spread and many outliers, indicating significant income variability. The Technical/Sales/Services category also shows a higher median income, though the distribution is somewhat narrower. The Other (production workers/farmers) group has a lower median income, with a spread similar to that of the Not working category, which has the lowest median income. This suggests that higher-paying jobs are generally found in managerial/professional and technical/sales/services occupations, while individuals in non-working or lower-skill jobs tend to have lower income.

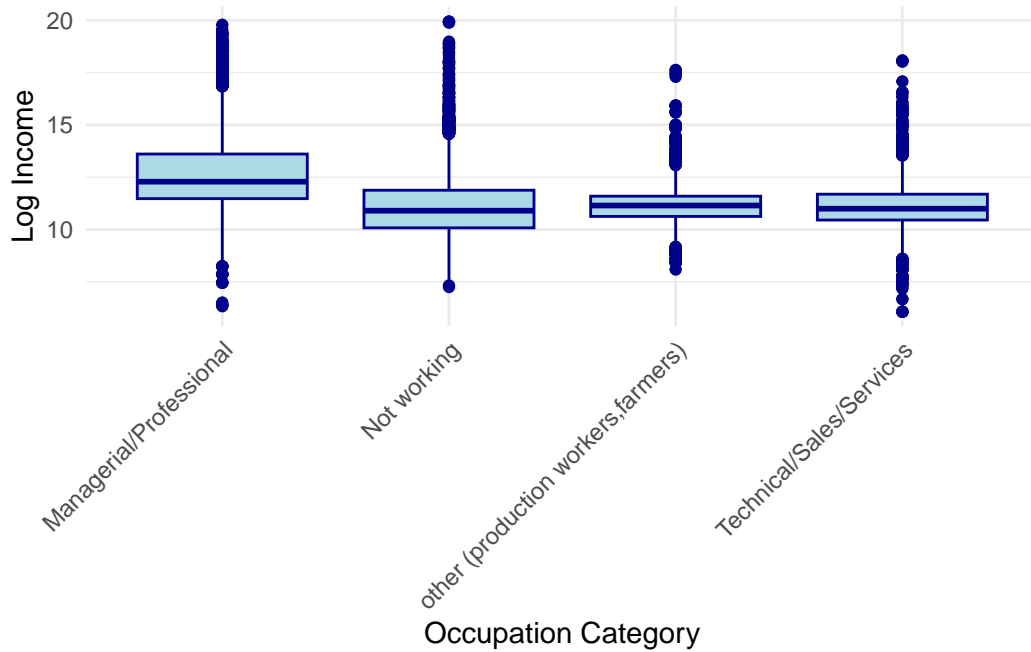


Figure 9: Log Income by Occupation Category

3 Model

The goal of our modelling strategy is twofold. Firstly, we aim to accurately predict the total income of individual in United States that participated in the survey of 2022. Secondly, we seek to evaluate the efficacy of different modeling approaches, from simple linear regression to Bayesian generalized linear model - to understand their predictive capabilities and assess the underlying relationships between variables. By comparing these models, we can determine which approach provides the most robust and reliable predictions, which considering the variability and potential uncertainty in this data.

Here we briefly describe the Bayesian analysis model used to investigate. Background details and diagnostics are included in Section [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](#).

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

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:

$$\begin{aligned}\widehat{\text{income}} &= \beta_0 + \beta_1 \cdot \text{Sexmale} + \beta_2 \cdot \text{Education Level} + \beta_3 \cdot \text{Marital Status} + \beta_4 \cdot \text{Labor Force} \\ &\quad + \beta_5 \cdot \text{Work Status} + \beta_6 \cdot \text{Occupation Category}\end{aligned}$$

From Figure [10](#), we can see that our linear model do not show a good accuracy for the Total Income variable. The graph shows presence of various outliers which causes the graph to be focused near the start of the range. From the red dotted line, we can see that it has show a slight predictive power but still weak in this graph.

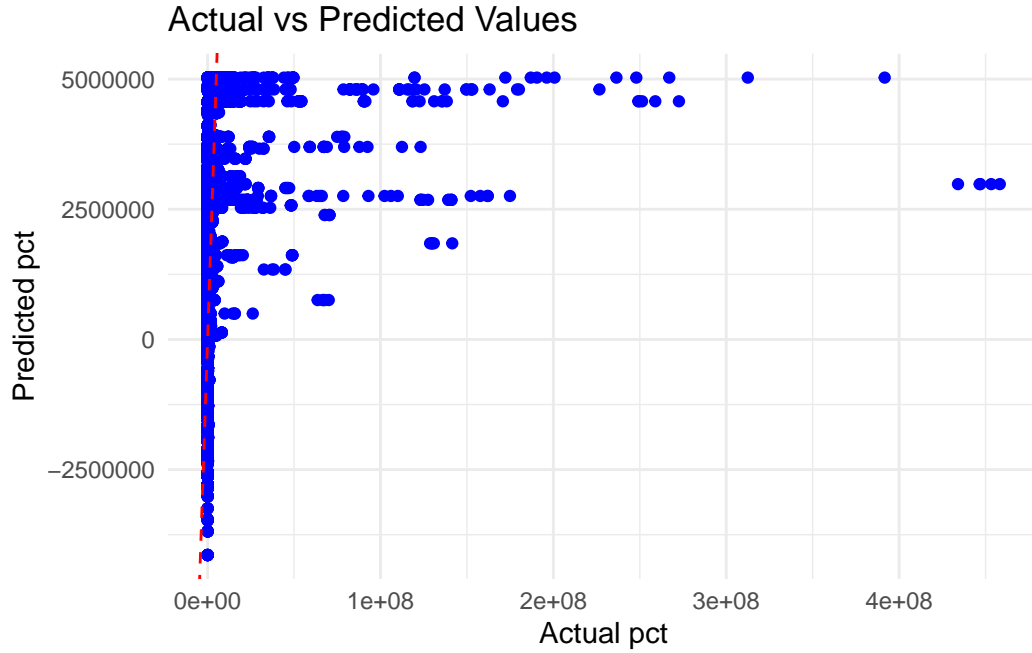


Figure 10: Linear Model has limited predictive accuracy for Total Income

Table 1

Variable	Estimate	Std. Error	t-value	Pr(> t)	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 Log Transformation of Linear Model

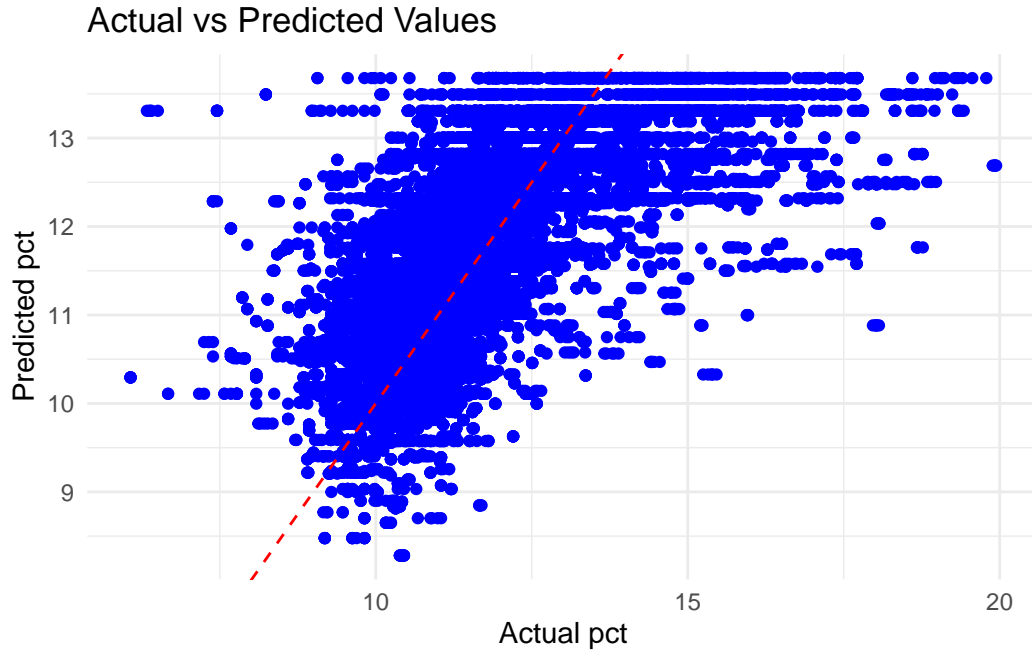


Figure 11: Log Linear Model has better predictive accuracy for log Total Income

Table 2

Variable	Estimate	Std. Error	t-value	Pr(> t)	Significance
(Intercept)	11.359653	0.089632	126.736	< 2e-16	***
Sexmale	0.421812	0.027118	15.555	< 2e-16	***
educ	0.184963	0.003427	53.97	< 2e-16	***
married	-1.01478	0.024175	-41.976	< 2e-16	***
lf	-0.122501	0.036421	-3.363	0.000771	**
num_work	0.488869	0.014911	32.786	< 2e-16	***
num_occu	-0.533513	0.012695	-42.025	< 2e-16	***
Residual Standard Error	1.273				
Multiple R-squared	0.4379				
Adjusted R-squared	0.4378				
F-statistic	2957				
p-value	< 2.2e-16				

Table for Log Linear Model Results

3.4 Bayes Model

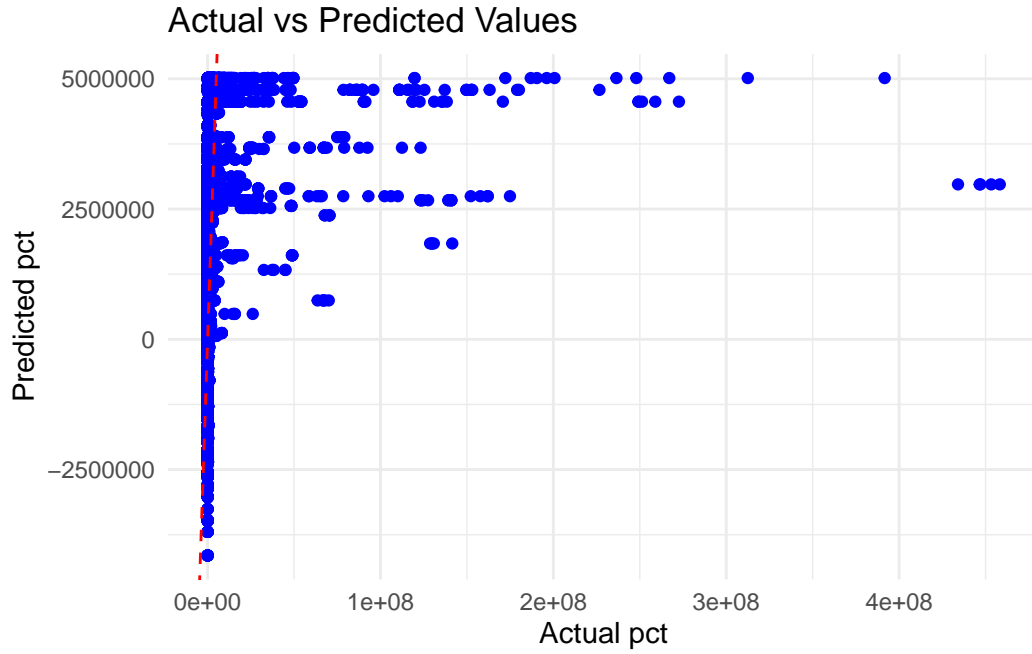


Figure 12: Bayes Model has limited predictive accuracy for Total Income

Table 3

Variable	Mean	SD	10%	50%	90%	mcse	Rhat	n_eff
(Intercept)	-330526.5	890630.7	-1458852.6	-334319.1	802368.2	NA	NA	NA
Sexmale	665201.5	264380.1	323626.9	663194.0	1003960.6	NA	NA	NA
Education Level	227389.2	33021.8	185831.1	226841.4	270562.7	NA	NA	NA
Marital Status	-1106314.8	236391.0	-1406586.4	-1110600.3	-803618.0	NA	NA	NA
Labor Force	109757.1	362967.1	-354290.6	110247.3	571555.5	NA	NA	NA
Work Status	1893452.3	143077.3	1714955.2	1891923.1	2076167.0	NA	NA	NA
Occupation Category	-1276316.9	124229.1	-1438524.9	-1274086.5	-1116421.2	NA	NA	NA
sigma	12358471.7	57949.8	12283774.9	12358169.3	12431827.0	NA	NA	NA
mean_PPD	1610015.7	113810.9	1461896.5	1609532.0	1755087.9	NA	NA	NA
mcse	14336.4	4097.4	467.9	3636.0	571555.5	1825.7	1	388
Rhat	1.0	1.0	1.0	1.0	1.0	NA	1	385
n_eff	1.0	1.0	1.0	1.0	1.0	NA	1	416

Table for Log Bayes Model Results

3.5 Bayes Model on Log Transformation

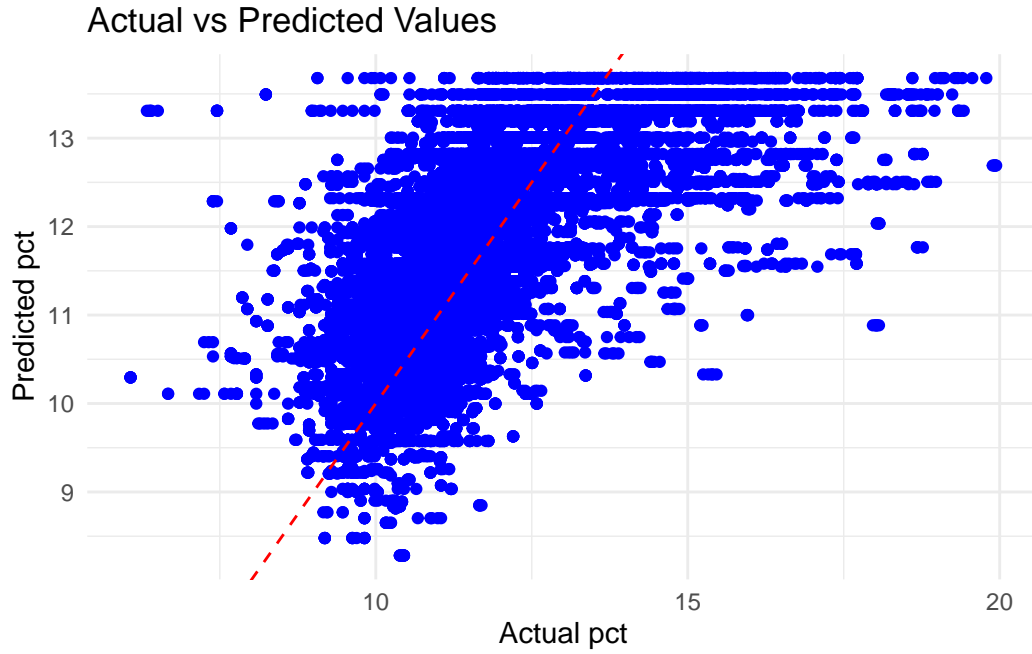


Figure 13: Bayes Model has better predictive accuracy for log Total Income

Table 4

Variable	Mean	SD	10%	50%	90%	mcse	Rhat	n_eff
(Intercept)	11.4	0.1	11.2	11.4	11.5	NA	NA	NA
Sexmale	0.4	0.0	0.4	0.4	0.5	NA	NA	NA
educ	0.2	0.0	0.2	0.2	0.2	NA	NA	NA
married	-1.0	0.0	-1.0	-1.0	-1.0	NA	NA	NA
lf	-0.1	0.0	-0.2	-0.1	-0.1	NA	NA	NA
num_work	0.5	0.0	0.5	0.5	0.5	NA	NA	NA
num_occu	-0.5	0.0	-0.5	-0.5	-0.5	NA	NA	NA
sigma	1.3	0.0	1.3	1.3	1.3	NA	NA	NA
mean_PPD	11.8	0.0	11.8	11.8	11.8	NA	NA	NA
mcse	0.0	0.0	0.0	0.0	0.0	0	1	3813
Rhat	1.0	1.0	1.0	1.0	1.0	0	1	3759
n_eff	1.0	1.0	1.0	1.0	1.0	0	1	4603

Table for Log Bayes Model Results

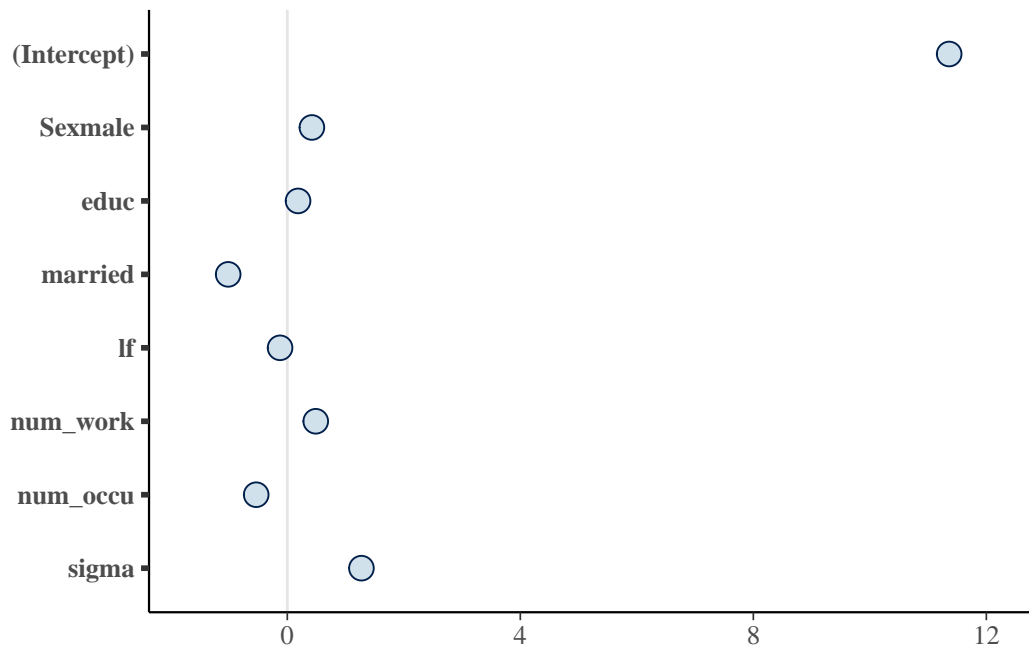
3.6 Model justification

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. In overall, the results from the Bayes log model show a comprehensive analysis of the relationship between various predictors and the log-transformed income. The estimates for the coefficients, such as `Sexmale` (0.4), `educ` (0.2), and `num_work` (0.5), indicate that being male, having more education, and working more hours are all associated with higher log income. Married individuals, however, have a negative coefficient (-1.0), suggesting that being married is associated with a decrease in log income, although this relationship is modest. The model also shows a sigma value of 1.3, indicating the residual standard error, which measures the spread of the log income data around the model's predictions. The mean PPD (11.8) suggests that the posterior predictive distribution for income is centered around this value, reinforcing the model's fit. The MCMC diagnostics (Rhat values of 1.0 for all parameters) confirm that the model has converged, with effective sample sizes (`n_eff`) indicating adequate sampling efficiency for each parameter. Lastly, the model appears to be well-specified, with clear relationships between income and the included variables, and good convergence as evidenced by the diagnostics.

4 Results

Our results are summarized in Table 4.

The graph displays the posterior estimates of the model parameters in a plot, with each circle representing the point estimate of a parameter and its corresponding value plotted along the x-axis. Most of the estimates are clustered near zero, indicating relatively small effects on the outcome variable, except for sigma, which is the largest estimate, located at the far right of the graph. This suggests that the model accounts for a substantial amount of variability in the dependent variable. The plot visually emphasizes the magnitude of the coefficients, with some parameters having minimal effect, while sigma represents the level of uncertainty in the model's predictions. The results from the graph suggest that the majority of the model's predictors, have relatively small point estimates, but these estimates are still substantial enough to be considered meaningful in the context of the model.



The model summary table shows key results from the regression analysis, with estimates for both the coefficients and model diagnostics. The Intercept is 11.36, indicating a significant baseline value for the outcome variable when all predictors are zero, with a small standard error (0.09), suggesting high precision. Sexmale has a coefficient of 0.42, statistically significant with a small standard error of 0.03, meaning that being male is associated with an increase in the outcome variable. The Education Level estimate (0.18) suggests a positive but modest effect on the outcome, with the very small standard error (0.00), indicating a highly precise estimate. Marital Status and Labor Force have negative coefficients of -1.01 and -0.12, respectively, suggesting that being married or part of the labor force is associated with a decrease in the outcome. Occupation Category also shows a negative relationship (-0.53), indicating a significant reduction in the outcome with different occupation categories. The model's R^2 of 0.438 indicates that about 44% of the variance in the dependent variable is explained by the model, while the Adjusted R^2 confirms this relationship is statistically valid, considering the number of predictors. Log.Lik. (-37811.782) and ELPD (-37819.9) are used for model comparison, with lower values indicating better model fit. The RMSE of 1.27 suggests the model predicts the outcome with reasonable accuracy, with a small average error in predictions.

Table 5: Explanatory models of factors influencing income

Table 5: Model Summary Table

Variable	Estimate	Std.Error
(Intercept)	11.360	0.09
Sexmale	0.420	0.03
Education Level	0.180	0
Marital Status	-1.010	0.02
Labor Force	-0.120	0.04
Work Status	0.490	0.01
Occupation Category	-0.530	0.01
Number of Observation	22778.000	
R2	0.438	
R2 Adj.	0.438	
Log.Lik.	-37811.782	
ELPD	-37819.900	
ELPD s.e.	179.700	
LOOIC	75639.700	
LOOIC s.e.	359.500	
WAIC	75639.700	
RMSE	1.270	

5 Discussion

5.1 Summary of Findings

In this paper, we have developed a Bayes linear model predicting the Total Income of United States individuals. The model has provided a accurate prediction in terms of middle income group. However, it is still hard to predict high and low income group. With our analysis, we have found that

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

B Model details

B.1 Posterior predictive check

In Figure 14a we implement a posterior predictive check. The plot displays the posterior predictive check for a Bayesian model. The dark line represents the observed data (denoted as y), while the lighter, faded lines represent the posterior predictive distribution (denoted as y_{rep}), which shows the distribution of replicated data based on the model's parameters. From the graph, we can observe that the posterior predictive distribution (y_{rep}) closely follows the observed data (y) distribution, especially around the peak between values 10 and 14. The model seems to capture the central tendency and variability of the observed data, as evidenced by the overlap of the two distributions. This suggests that the model is performing well in terms of replicating the observed data. However, the wider spread of the y_{rep} lines at higher values suggests some variability or uncertainty in the model's predictions, which is expected in Bayesian modeling.

In Figure 14b we compare the posterior with the prior. The graph presents posterior and prior distributions for the parameters of a Bayesian model. On the left, the posterior distributions show the estimated values of each parameter based on the data and the model. The vertical lines represent the credible intervals for each parameter, and the dots represent the point estimates. On the right, the prior distributions are shown for the same parameters, illustrating what the model assumed about each parameter before observing the data. The prior distributions are much wider and more diffuse compared to the posterior distributions, indicating less certainty about the parameters before incorporating data. The priors for most parameters have a wide spread, with the prior for Intercept being particularly broad, suggesting high uncertainty. The difference between the prior and posterior for each parameter shows the influence of the data in updating the beliefs about these parameters. This plot provides a useful comparison to see how the data has informed the model's parameter estimates.

B.2 Diagnostics

Figure 15a is a trace plot. The graph displays trace plots for multiple parameters in a Bayesian model, generated through Markov Chain Monte Carlo (MCMC) sampling. The plot contains four separate MCMC chains, each representing an independent sequence of parameter samples, with the x-axis indicating the number of iterations (from 1 to 1000) and the y-axis showing the values of the parameter estimates at each iteration. The trace plots for each parameter indicate good convergence across all chains, with the chains mixing well and stabilizing at similar values over time. The lack of systematic drift and the stable fluctuations around central values suggest that the model has reached stationarity, and the sampling process is not biased.

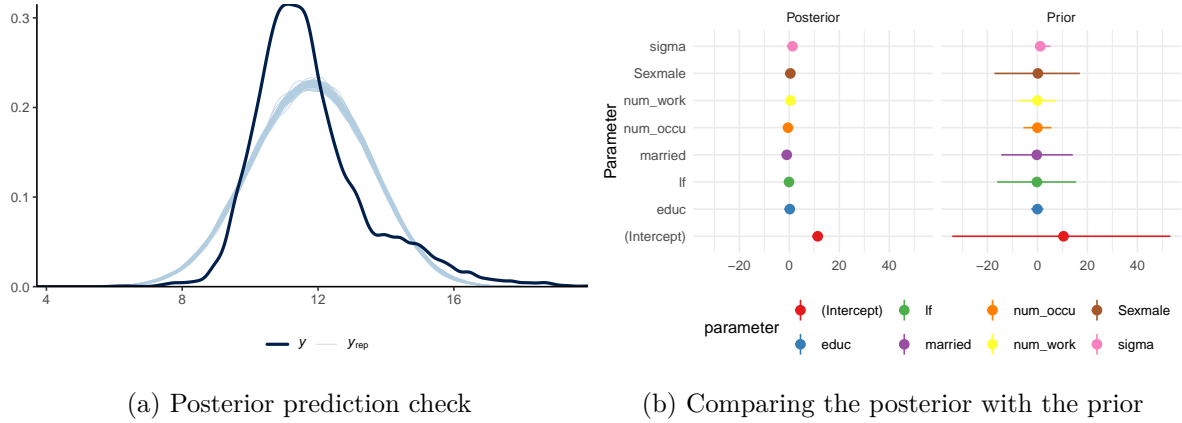


Figure 14: Examining how the model fits, and is affected by, the data

by the starting values. This indicates that the MCMC chains are successfully exploring the posterior distribution, and the parameter estimates are reliable. Overall, the graph suggests that the model's parameters are well-estimated, with the chains converging to a consistent distribution, making the results trustworthy for our models.

Figure 15b is a Rhat plot. This plot represents the Gelman-Rubin diagnostic (denoted as \hat{R}), which is used to assess convergence in a Bayesian analysis. The x-axis shows the values of \hat{R} , which indicates whether the multiple MCMC chains have converged to the same distribution. Most of the points lie very close to 1.00, indicating that these parameters have likely converged, as the chains are sampling from the same posterior distribution. A few points may be slightly above 1.00, but all are still well below 1.1, suggesting that the model is close to convergence. The plot shows $\hat{R} \approx 1.05$ for several parameters, which is generally considered acceptable for convergence, while any parameters with \hat{R} above 1.1 would indicate poor convergence, but this is not visible in the plot. The small variation in \hat{R} values suggests the model is performing well and the chains have likely mixed adequately.

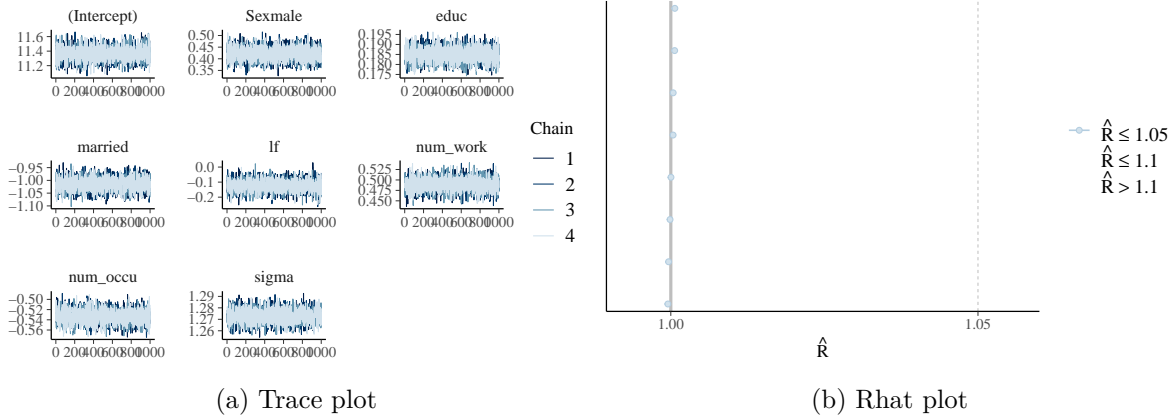


Figure 15: Checking the convergence of the MCMC algorithm

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>.
- Developers, Apache Arrow. 2022. “Apache Arrow: A Cross-Language Development Platform for in-Memory Data.” <https://arrow.apache.org/>.
- 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/>.
- Wickham, Hadley et al. 2022a. “Dplyr: A Grammar of Data Manipulation.” <https://dplyr.tidyverse.org/>.
- et al. 2022b. “Tibble: Simple Data Frames.” <https://tibble.tidyverse.org/>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolmund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wu, Haiyan. 2022. “kableExtra: Construct Complex Table with ‘Kable’ and ‘Markdown’.” <https://haozhu233.github.io/kableExtra/>.
- Xie, Yihui. 2022. “Knitr: A General-Purpose Dynamic Report Generation Tool.” <https://yihui.org/knitr/>.