STA302 Fall 2023 Methods of Data Analysis 1
Final Project Report- Part 3

Names of Group Members	Contribution
Dian Rong	Results
Krishna Kumar	Introduction
Carter Ponce	Discussion + Ethics
Jing Yu	Methods

INTRODUCTION

Within the landscape of socioeconomic factors shaping individual well-being, our research focuses on analyzing the relationship between lifestyle elements and an individual's family income-to-poverty ratio in the United States from 2017 to 2018. The numerical predictors are alcohol consumption, hours slept, and BMI while the categorical predictors are marital status and smoking habits.

This investigation stems from recognizing the profound impact of BMI, marital status, sleep patterns, alcohol consumption, and smoking habits on financial standing, offering valuable insights for policymakers, healthcare professionals, and individuals alike. The impact of these variables will be determined by their coefficients in the linear model: the higher the coefficient is, the bigger the impact the variable has.

Understanding these relationships is crucial for informed decision-making, resource allocation, and societal progress. Past studies regarding this subject have been conducted and have concluded fascinating results: One study concluded that moderate-to-heavy alcohol drinkers have higher household incomes than non-drinkers and individuals who never drank (Ormond & Murphy, 2016). Another study found that married households exhibit the highest monthly income, with widowed females having the lowest (Dunga, 2017). A third study found that family income correlates with obesity prevalence differently for men and women: for women, the prevalence of obesity decreased as income increased, and for men, the prevalence of obesity was lower in the lowest and highest income groups (Ogden et al., 2017).

While antecedent studies have begun researching this subject, our work stands apart by comprehensively examining the impact of BMI, marital status, sleep, alcohol, and smoking within the context of individuals aged 30 years or older in 2017-2018.

Our work fills the gap in past studies by understanding the significance of these variables and how they individually impact the income-to-poverty ratio of a given individual's family.

The study objective is to examine and quantify the impact of lifestyle factors on an individual's family income-to-poverty ratio through rigorous statistical analysis. The research aims to provide nuanced insights into the complex nature of these variables and their influence on socioeconomic dynamics during the specified time period.

METHODS

We examined data from the US National Health and Nutrition Examination Survey (NHANES), specifically focusing on 2017 to 2018 data for individuals aged 30 or older. Merging datasets, we gathered variables including BMI, MaritalStatus, sleep hours, average alcohol intake, smoking habits, and the response variable, INDPMPIR (income-to-poverty ratio). In the preprocessing step, we removed missing values and encoded variables to suitable forms.

Our initial model encompassed three numerical predictors (BMI, sleep hours, alcohol intake) and two categorical predictors (MaritalStatus, smoking habits) with INDPMPIR as the response variable. To ensure residual plots would be reliable, first we verified two conditions in MLR: conditional mean response and conditional mean predictor. We checked the conditional mean response by examining scatterplots of response against fitted values for random diagonal scatter or easily identifiable non-linear trends. Simultaneously, pairwise scatterplots of numerical predictors were assessed for linear relationships or no patterns. Meeting these criteria ensured reliable residual plots; otherwise, interpretations could be misleading.

After establishing the essential conditions for our MLR model, we verified its assumptions, crucial for our analysis' accuracy. Specifically, we examined scatterplots for residuals against fitted values and each numerical predictor, and used boxplots for categorical predictors. To identify violations, we looked for systematic patterns or clustering indicating uncorrelated error violations. We also checked for non-linearity by observing curve-like patterns in residuals and assessed variance constancy by noting spreading patterns. A QQ plot assessed normality assumptions, focusing on stark deviations from the diagonal line. Our preliminary model only had violation in normality, so we addressed it by using the Yeo-Johnson Power transformation on the response.We used this type of transformation instead of Box-Cox because there were many zero values which cannot be handled by Box-Cox. After correcting the initial model, we plotted the graphs displaying the assumptions and conditions again to ensure issues were addressed, compared with the previous one.

Additionally, we conducted model diagnostics to spot problematic observations and assess multicollinearity among predictors. Establishing cutoffs, we calculated various measures for leverage, outliers, and all three influential observations for both untransformed and transformed models. Comparing the number of problematic observations between different models aided in identifying transformations' effectiveness. We refrained from removing problematic observations unless contextually warranted, aiming for generalizability. To evaluate multicollinearity, we used the variance inflation factor (VIF), setting a cutoff at 5 for severe cases. In the absence of significant multicollinearity, we explored model selection to identify potential improvements.

In our model selection, we used both manual and automated methods. Initially, we tried all possible subsets, evaluating models using four numerical criteria: adjusted R-squared, AIC, corrected AIC, and BIC. Then, we explored three automated methods (forward, backward, stepwise) with AIC and BIC to generate other models. Comparing these models based on the same four criteria, alongside considerations of multicollinearity, problematic observations, and model assumptions, we pinpointed the most suitable model.

RESULTS

Based on interest and literature, we decided to focus on the effect of average alcohol intake, smoking habits, BMI, hours of sleep, and marital status on INDPMPIR. After all, an individual's smoking habits can lead to health problems, absenteeism from work, and decreased job performance, all combining to decrease income. This negative correlation has been found by Casetta et al. in 2017. On the other hand, other vices such as alcohol, can be indicative of wealth, and associated with higher income (Ormond & Murphy, 2016). BMI is more directly associated with health and quality of living, as individuals with certain BMIs may experience health issues causing different employment opportunities and work productivity. This is supported by Ogden et al., 2017 who detected a higher prevalence of obesity in lower income households. Another health-related factor is average sleep duration, which can affect well-being and energy. Longer sleep has been shown to improve productivity and therefore wages (Gibson & Shrader, 2014). Similarly, specific marital statuses can also offer socioeconomic benefits such as pooling income and sharing living expenses, and Steven Dunga in 2017 concluded that married households have the highest monthly income, compared to other households who're widowed, divorced, and single.

Table 1: Summary Statistics for our numerical variables

Variable Name	Min.	Median	Mean	Max	SD	IQR	Skewness	Outliers
BMI	14.9	29.0	30.31	86.2	7.37	8.475	1.351	81
SleepHrsNight	2.0	7.5	7.496	14.0	1.53	2.0	-0.06618	39
AlcoholDay	0.0	2.0	2.34	15.0	1.94	2.0	2.803	84

Minimum, median, mean, maximum, standard deviation, interquartile range, skewness, and number of outliers for each of BMI, average hours of sleep every night, and average drinks consumed when drinking

Table 2: Distribution of Marital Status Categories

Marital Status	Total
Divorced	341
Live with their Partner	202
Married	1308
Never Married	258
Separated	85
Widowed	156

Most individual surveyed are married, with the other categories having similar, small totals

Table 3: Distribution of Smokers and Non-Smokers

Currently Smokes	Total
No	1231
Yes	1119

The distribution of smokers and non-smokers in our Data is very similar

As seen in Table 1, BMI and AlcoholDay have a slight and strong right skew respectively. The table also shows all numerical values have extreme outliers that may be nonsensical, such as a BMI of 86.4, 15 alcoholic drinks in a day when drinking, and 2 hours of sleep a night.

Table 2 shows the vast majority of individuals surveyed are married, while the other categories are fairly even. Similarly, the number of individuals who smoke and don't smoke are very similar (Table 3).

Our initial model looked at the effect of all the variables in Table 1-3 on INDPMPIR. However, when checking our model assumptions, we noticed that the QQ plot didn't have a diagonal line and therefore the model failed the normality assumption. Then, we used the Yeo-Johnson power transformation which estimated the power with the highest log-likelihood as 0.5 (Figure A1). We then created a transformed model which contains the same predictors as the initial model, but the response has now been square rooted. The transformed model passes all assumptions and conditions.

Then, in an attempt to find a better model, we used automated and manual selection methods from the transformed model. Specifically we used six automated methods: one using AIC and one using BIC for each direction (forward, backward, or stepwise). Ultimately, each method using AIC returned the same model and each method using BIC returned the same model. The AIC methods returned the original transformed model, while the BIC methods returned a new model. This BIC model had TransformedIncomeRatio as the response variable and marital status, smoking status, average alcohol consumption, and average sleep duration as the predictors. For manual selection, we used the leaps library and the all possible subsets method to find the best model with each number of predictors from one to four. This process highlighted three new models as the best model with four predictors was the same as the one obtained from the BIC methods. All the new models from model selection satisfied the model assumptions and conditions so were eligible to be our best model.

Table 4: Notable Total Problematic Observations and Correctness Measures for Each Model

Model	Average Number of Influential Points on Each Coefficient	R^2	AIC	Corrected AIC	BIC
Transformed Model	3.666666667	**0.1300072**	**-3262.192**	-3204.389	**-3262.081**
BIC Selected Model	4.125	0.1282528	-3258.367	**-3206.344**	-3258.275
1 Predictor Model	**3.2**	0.09171743	-3163.11	-3128.429	-3163.064
2 Predictor Model	4.333333333	0.1140744	-3221.753	-3181.291	-3221.692
3 Predictor Model	3.285714286	0.1222397	-3242.914	-3196.671	-3242.838

The average number of influential points on each coefficient and the value of correctness measures for each model is displayed above. The best option in each column is bolded, including the best model which is the transformed model.

The models were compared in three main aspects: multicollinearity, number of problematic observations and values of correctness measures. All models had low multicollinearity (VIF values <1.1), zero leverage points, zero outlier points, zero points that were influential on all fitted values, and zero points that were influential on their own fitted value. As seen in Table 4, the best model was the transformed model in all correctness measures except corrected AIC. Additionally, the transformed model had a similar average number of influential points on each coefficient as the model with the smallest average. Therefore we ultimately prefer the transformed model.

DISCUSSION

Table 5: Coefficient Values of the Final Model

	Value	Pr(>|t|)
Intercept	1.869558	<2e-16
MaritalStatusLivePartner	0.002422	0.9562
MaritalStatusMarried	0.297455	<2e-16
MaritalStatusNeverMarried	-0.037533	0.3611
MaritalStatusSeparated	-0.054934	0.3644
MaritalStatusWidowed	0.091357	0.0584
BMI	-0.003394	0.0160
SleepHrsNight	-0.028915	2.03e-05
AlcoholDay	-0.027148	6.73e-07
SmokeNowYes	-0.143291	1.55e-11

Information about each coefficient in the final model including the value output for relevant figures relating to final model coefficients, variables, and intercepts.

The chosen linear model, with TransformedIncomeRatio as the response variable and predictors including MaritalStatus, BMI, SleepHrsNight, AlcoholDay, and SmokeNow, offers valuable insights into the relationship between lifestyle factors and an individual's family income-to-poverty ratio. The coefficients provide information on the strength and direction of these relationships. For instance, being married (MaritalStatusMarried) is associated with a positive impact on income-to-poverty ratio, while higher BMI, lower sleep duration, and alcohol consumption have a negative association. Smoking (SmokeNowYes) also negatively impacts income-to-poverty ratio. The model's overall fit is summarized with the following: multiple R-squared is 0.1333, adjusted R-squared is 0.13, F-statistic is 40.72 on 9 and 2383 degrees of freedom and the p-value is < 2.2e-16.

The research question focused on examining and quantifying the impact of lifestyle factors on an individual's family income-to-poverty ratio. The model allows us to explicitly answer this question by interpreting the coefficients. For example, being married and having higher sleep hours positively influence the income-to-poverty ratio, while higher BMI, alcohol consumption, and smoking have negative effects.

The findings align with some existing literature. For instance, the positive impact of marriage on income echoes Steven Dunga's conclusion that married households tend to have higher monthly incomes. However, the negative impact of smoking on income contradicts the study by Casetta et al., which found a negative correlation between smoking habits and income. These inconsistencies highlight the complexity of socioeconomic factors and the need for nuanced analysis.

While the final model addresses the initial issue of non-normality through the Yeo-Johnson transformation, some issues persist. Notably, the p-value for MaritalStatusNeverMarried is relatively high (0.3611), suggesting non-significance. The impact of this issue is twofold. Firstly, the variable might not be contributing significantly to the model, potentially affecting its overall explanatory power. Secondly, it raises questions about the generalizability of the findings to never-married individuals.

Addressing the issue of non-significance for MaritalStatusNeverMarried might involve reconsidering the inclusion of this variable in the model. However, given the research objective to comprehensively examine the impact of all chosen variables, and considering the lack of severe multicollinearity or other violations, retaining the variable is justified. Its potential contribution, even if non-significant, adds to the completeness of the analysis.

In conclusion, the final model provides meaningful insights into the relationship between lifestyle factors and family income-to-poverty ratio, contributing to the existing body of knowledge. Lingering issues, while acknowledged, are deemed acceptable for the sake of a comprehensive analysis. The findings underscore the intricate effect of socioeconomic factors and the importance of considering various lifestyle elements when assessing financial well-being.

ETHICS

Adhering to ethical standards is paramount to the integrity of research. Our study, utilizing US National Health and Nutrition Examination Survey (NHANES) data, aligns with ethical guidelines, ensuring informed consent and safeguarding individual privacy.

In preprocessing, we responsibly managed missing values and encoded variables, emphasizing accuracy in analysis. Ethical transparency dictated the acknowledgment of outliers and extreme values in result reporting, fostering truthful and unbiased presentation.

Opting for both manual and automated selection methods was a deliberate ethical choice. This approach addresses concerns related to responsibility for potential harms, as it combines the efficiency of automation with the oversight and nuanced judgment of manual selection. Our commitment to avoiding negligence or recklessness is reflected in this dual-method strategy.

While automated tools streamline processes, they may lack the depth of ethical consideration inherent in manual selection. By employing both methods, we aimed for a balanced approach, preventing the blind transfer of responsibility to automated tools.

Our ethical stance extends beyond statistical methods, emphasizing accountability in model selection choices. The combined approach reflects a commitment to robust and unbiased results, underlining our dedication to ethical research practices. In the complex landscape of statistical ethics, our dual-method approach represents a conscious effort to uphold ethical standards while navigating the intricacies of model development.

REFERENCES

Casetta, B., Videla, A. J., Bardach, A., Morello, P., Soto, N., Lee, K., ... & Ciapponi, A. (2017). Association between cigarette smoking prevalence and income level: a systematic review and meta-analysis. *Nicotine & Tobacco Research, 19*(12), 1401-1407.

Dunga, S. H. (2017). A gender and marital status analysis of household income in a low-Income township. *Studia Universitatis Babes Bolyai-Oeconomica, 62*(1), 20-30.

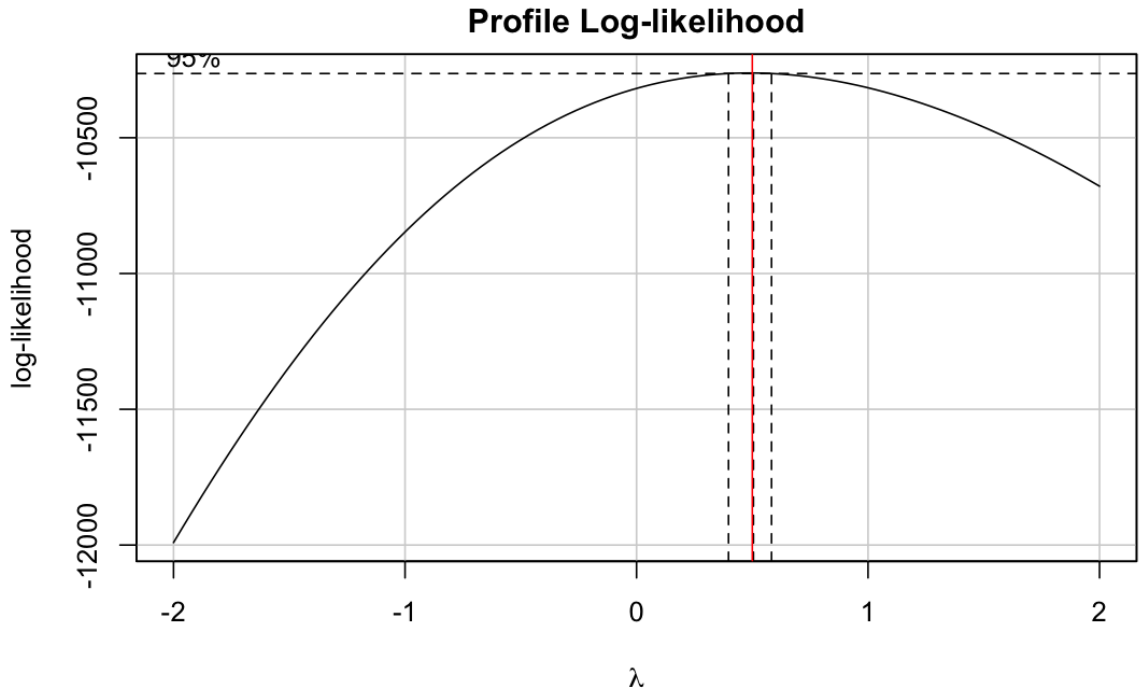
Gibson, M., & Shrader, J. (2014). Time Use and Productivity: The Wage Returns to Sleep. *UC San Diego: Department of Economics, UCSD*. Retrieved from https://escholarship.org/uc/item/8zp518hc

Ogden, C. L., Fakhouri, T. H., Carroll, M. D., Hales, C. M., Fryar, C. D., Li, X., & Freedman, D. S. (2017). Prevalence of obesity among adults, by household income and education—United States, 2011–2014. *Morbidity and Mortality Weekly Report*, 66(50), 1369.

Ormond, G., & Murphy, R. (2016). The effect of alcohol consumption on household income in Ireland. *Alcohol*, 56, 39-49.

APPENDIX

Figure A1: The Log-Likelihood of Various Yeo-Johnson Power Transformations



Of all possible Yeo-Johnson Power Transformations, λ=0.5 as indicated with a red line, has the highest log-likelihood at around 95%