Analysing how health and lifestyle factors effect individual Income level*

Including BMI, Sleep hours, depression degree, year of drinking alcohol, and age

Ruiying Li & Lorina Yang

2024/10/03

Contribution

Name: Hanqing(Lorina) Yang Contribution: responsible for writing introduction, ethics discussion and part of preliminary result evaluation. Name: Ruiying Li Contribution: responsible for writing data description, coding part and evaluating model result.

Introduction

Household income plays a central role in shaping an individual's quality of life, purchasing power, and ability to achieve personal and professional goals. Simultaneously, health factors such as obesity, depression, alcohol consumption, and sleep patterns have significant impacts on both physical well-being and economic outcomes. The aim of this study is to investigate whether a linear relationship exists between the household income-poverty index and predictors such as BMI, depression levels, alcohol consumption, sleep hours, and age.

By analyzing these variables, we want to explore how personal health and lifestyle factors influence household income-poverty status. It is important because the findings could help guide policies addressing income-related health disparities, and create more targeted interventions, such as mental health support or obesity prevention programs, which may reduce healthcare costs and improve overall economic productivity.

The relationship between income and health has been widely studied. For example, a meta-analysis by Kim, T.J. found that lower income is strongly associated with subsequent obesity across Canada, the USA, and the UK, highlighting the link between financial status and health outcomes. Another study by Ridley et al. (2020) emphasized the bidirectional relationship between income and mental health, showing that mental illness can reduce employment and income, while financial hardships exacerbate mental health conditions like depression and anxiety. This underscores the importance of economic stability in maintaining mental well-being. Similarly, research on alcohol consumption by Cerdá, M. (2011) found that lower income is associated with both higher odds of abstinence and heavy drinking, compared to moderate drinking.

Given these findings, this study seeks to determine whether there is a linear relationship between household income-poverty ratio and health-related predictors. Linear regression is an appropriate tool for 2 reasons. First, it is effective in identifying trends and associations of the continuous responses variable and allows us to quantify the relationship between the income-poverty index (the dependent variable) and multiple predictors. Second, it also allows for the inclusion of multiple variables simultaneously, enabling control for confounding factors like age or alcohol consumption, which ensures more accurate and reliable results.

*Cleaning data resources available at https://github.com/Liruiying0414/cleaning-data

Data Description

The data set used in the proposal was founded in US governmental website called 'National Center for health statistics'(Pruim 2015) as an open data resources, and author who arrange data set called Randall Pruim, and the data set was collected by Michelle Dalrymple of Cashmere High School and Chris Wild of the University of Auckland, for use in teaching statistics, the collection methods were multiple,for the proposal data set, it is mainly concentrated on health screenings and self-reported surveys.

The origin data set contains 75 variables from sample year 2009 to 2012, concluding 10000 observations after re-sampling,including demographic variables such as Survey year, age, gender, race, Physical measurement like BMI, weight, Height, Health variables like Depressed, Diabetes, and Lifestyle variables like general health condition, number of drinking alcohol year etc. The original purpose for the data set was to examine the health and nutrition information for different races in the US, for an education purpose.

In this proposal,only 6 variables be selected, which is Age,Poverty,Depressed,Sleeping hours,BMI and alcohol year.

This proposal indicates Poverty as response, because it is a continuous variable that meets the requirement for building a linear regression model, and it represents a annual ratio of family income to poverty guidelines in percentage,smaller numbers indicate more poverty,and rest of 5 variables are predictors,exploring the linear relations between poverty level and these health and lifestyle factors.

BMI: A numeric variable represents Body mass index (weight/height2 in kg/m2),for participants aged over 18, measure obesity in our research question.

Age: A numeric variable indicates age in years at screening of study participant who over 18

Depressed: A categorical variable indicates self-reported number of days where participant who aged over 18 felt down, depressed or hopeless. (One of None, Several, Majority (more than half the days), or Almost All)

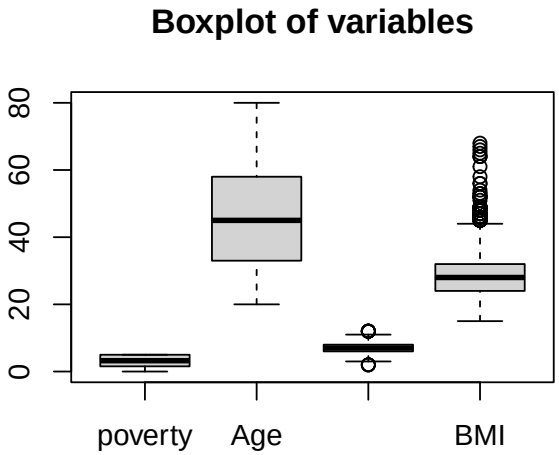
SleepHrsNight: A numeric variable represents self-reported number of hours study participant aged over 18 usually gets at night on weekdays or workdays.

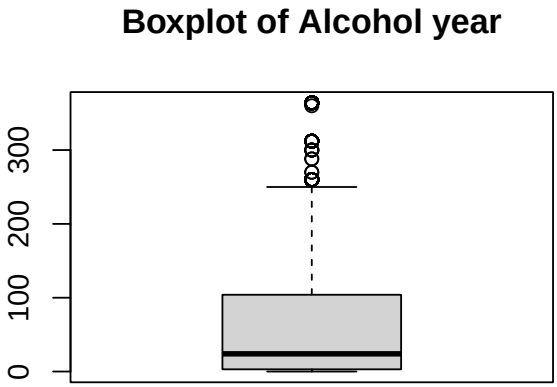
AlcoholYear: A numeric variable represents estimated number of days over the past year that participant aged over 18 drank alcoholic beverages

Those predictors assign numerical and categorical value for measure the health factors we intend to investigate, including obesity measure index BMI, age, depressed level, sleep hour and alcohol drinking days per years.

Here is the brief data set for proposal

```
##   X SurveyYr Poverty Age Depressed SleepHrsNight BMI AlcoholYear
## 1 1  2009_10    0.81  20      None            11  20           3
## 2 2  2009_10    0.85  20   Several             5  23           5
## 3 3  2009_10    0.38  20      None             8  23          60
## 4 4  2009_10    0.61  20      None             8  37           6
## 5 5  2009_10    1.66  20      None             8  24         260
## 6 6  2009_10    1.66  20      None             8  24         260
```





Through cleaning data,we found there is several missing values in variables Poverty, Depressed,and Sleep Hours.After building box plot, we found there are no outlier in general, except for variable BMI and Alcohol year, this is because many participants did not drink alcohol during observated year, while there still exist a group of participant strongly addicted with Alcohol, almost drinking everyday. Apart from outliers,we discover variable Alcohol year shows an strong trend of right skewness, while other shows nearly symmetric shape.

Multiple linear model for poverty level

```
##
## Call:
## lm(formula = Poverty ~ BMI + Age + Depressed + SleepHrsNight +
##     AlcoholYear, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8106 -1.3831  0.1515  1.4463  3.3509
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.1575255  0.2532519   4.571 5.08e-06 ***
## BMI              -0.0076651  0.0046154  -1.661 0.096875 .
## Age               0.0157945  0.0018582   8.500  < 2e-16 ***
## DepressedNone     0.8347178  0.1332512   6.264 4.35e-10 ***
## DepressedSeveral  0.3731257  0.1484607   2.513 0.012019 *
## SleepHrsNight     0.0777272  0.0225825   3.442 0.000586 ***
## AlcoholYear       0.0020881  0.0002937   7.110 1.48e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.565 on 2702 degrees of freedom
## Multiple R-squared:  0.08448,    Adjusted R-squared:  0.08244
## F-statistic: 41.55 on 6 and 2702 DF,  p-value: < 2.2e-16
```

For individuals in the “Most” depressed group, when BMI, Age, Sleep Hours, and Alcohol consumption are all zero, the average poverty index is 1.1575255. In this “Most” depressed group, holding all else constant:

Each one-unit increase in BMI is linked to a 0.0076651 decrease in the poverty index.

Each additional year of age is associated with a 0.0157945 increase in the poverty index.

Each extra hour of sleep per night increases the poverty index by 0.0777272.

Each additional unit of annual alcohol consumption increases the poverty index by 0.0020881.

Compared to the "Most" depressed group: The "Several" depressed group has a poverty index that is, on average, 0.3731257 higher.

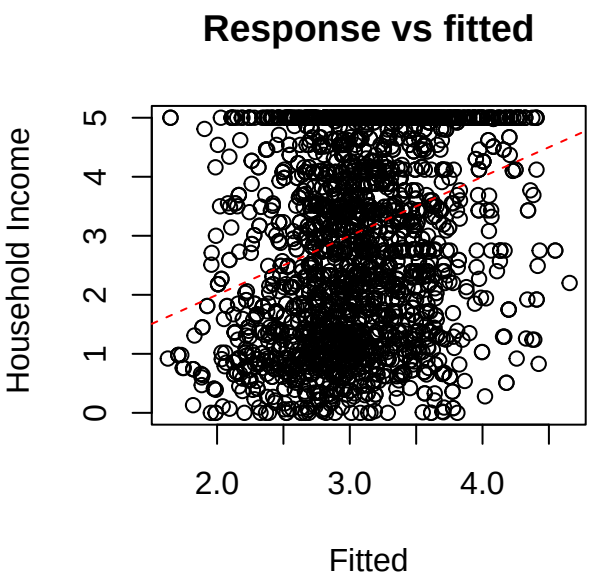
The "None" depressed group has a poverty index that is 0.8347178 higher.

Ethics Discussion

The dataset consisting of 20,293 observations on health factors and background, is considered trustworthy. It was collected by the U.S. National Center for Health Statistics, a credible source, along with widespread use and popularity ensure the reliability. However, ethical concerns remain, particularly regarding privacy and the impact on vulnerable groups. One issue is how statistical analysis could affect groups of stakeholders like the obese or low-income populations, potentially leading to discrimination and policy changes. Another issue is that some questions in the survey appear to be offensive and sensitive such as asking number of sex partners, however it is not that problematic since confidentiality and consent is be taken. Despite these concerns, the data set has been anonymized as each participant were given an ID instead of using real names, addressing confidentiality and privacy. Moreover, since the data was collected through a survey, which ensure the participants are volunteered to participate the study, and the questions design was obvious for participants to understand the aim of the survey, ensuring informed consent was obtained.

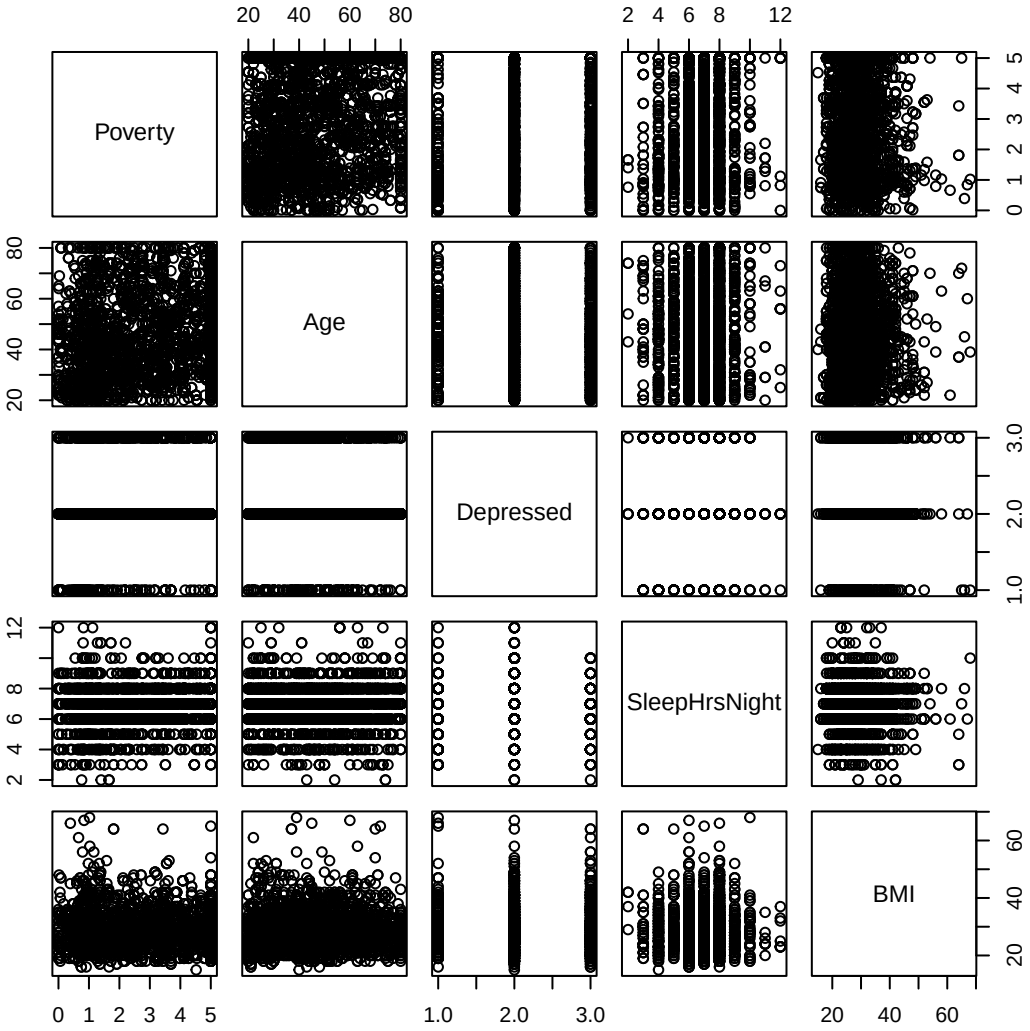
Preliminary Model Results

check linear regression condition 1



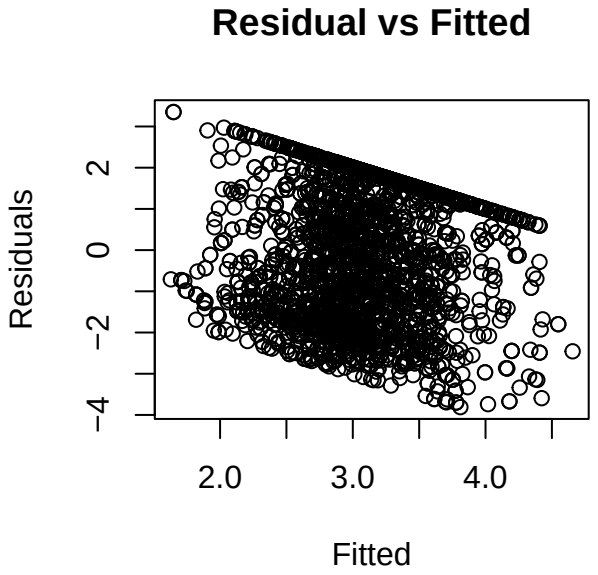
The plot about check condition 1 "Conditional mean response condition" for Multiple Linear Model. The condition 1 hold since the response-fitted plot show some easily identifiable non-linear pattern, indicate the mean responses "poverty" are a single function of a linear combination involving Beta parameter.

check condition 2



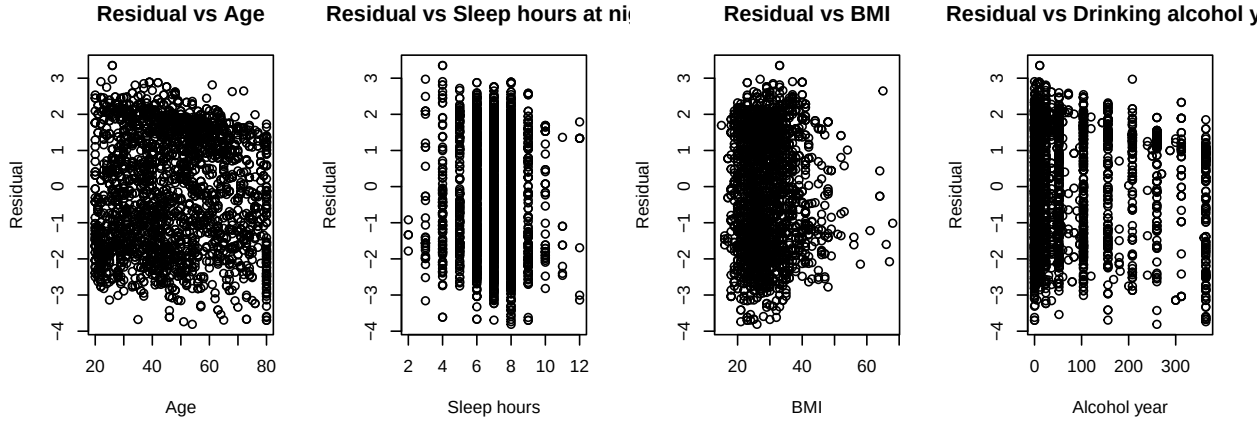
The pairwise scatterplots of predictors above check the condition 2 "Conditional mean predictor condition". The condition 2 hold since there is no curve and show non-linear pattern.Thus, the mean of each predictor is related to each other predictor in no more complicated way than linearly.

Residual vs fitted



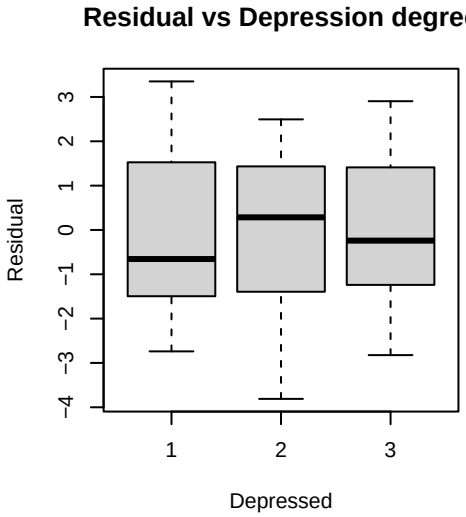
The Residual-Fitted plot violate linearity assumption because there exists a non-random pattern.

residual vs numeric predictors



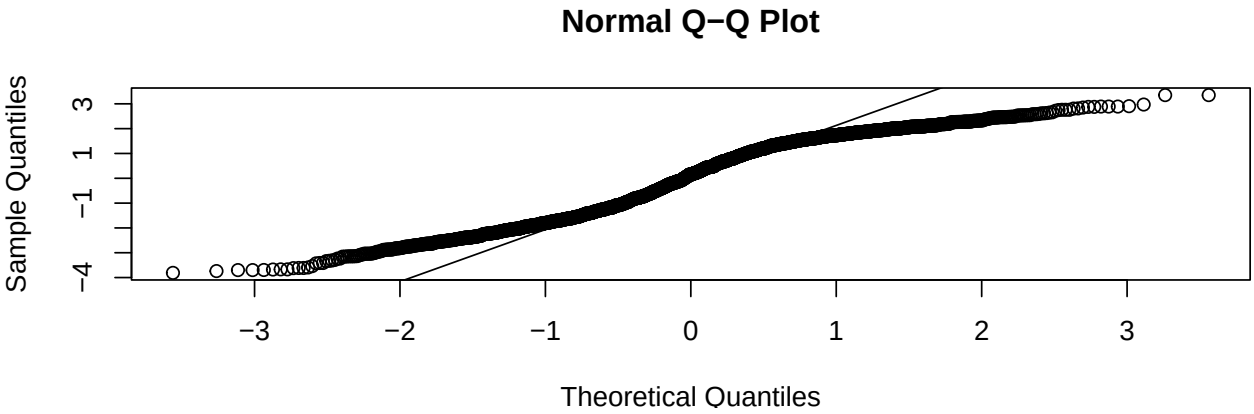
The residual-predictor plot show a random pattern, thus not violate any assumptions.

Residual vs categorical predictors



The Residual-Depressed predictor plot violates constant variance slightly, when looking at its boxplot for several depressed group compared to other two,the box is not in the same size.

use qqplot to check normality



From its qqplot of residual vs fitted, we could tell it also hold normality assumption, because its data points are close to the straight line in general and only few of them spread out.

Result summary

The linear regression sample in the proposal shows an random pattern in general,where 2 conditions and three assumptions are hold,except for some specific variables such as the boxplot for depressed level, which violates the constant variance assumption, and the residual vs fitted plot does not match the linearity assumption perfectly.

According to the linear regression model,The residual-fitted plot, residual-predictors plot, and QQ plot suggest that the multiple linear regression model is an appropriate tool for analyzing the relationship between poverty (measured by the income-poverty ratio) and predictors such as age, BMI, depression level, sleep hours, and alcohol consumption. However, some minor violations of linearity and constant variance assumptions were observed.

The results show a negative relationship between the income-poverty ratio and BMI (coefficient = -0.0077), supporting previous studies conducted by Kim, T.J.'s which indicated that individuals with higher BMI are more likely to live in poverty, notice lower income-poverty ratio implies higher poverty.

Interestingly, the model shows a positive relationship between alcohol consumption (measured by drinking days per year) and income-poverty ratio, suggesting that lower alcohol use is associated with higher poverty. This finding contradicts Cerdá, M.'s study, which found lower income linked to heavy drinking. However, this discrepancy may be due to other factors, as Cerdá's highlighted the association between lifetime income patterns and alcohol use decreased once we controlled for past-year income, education and occupation.

Additionally, the results show that individuals with higher depression levels are more likely to experience poverty, aligning with previous literature that links mental illness to reduced employment and income.

Other predictors, such as age and sleep hours, also show a positive relationship with the income-poverty ratio. Older individuals tend to have higher incomes, likely due to greater work experience, while those who sleep more are associated with healthier lifestyles, which may contribute to higher income and reduced poverty.

References

Pruim (2015) R Core Team (2024) Wickham et al. (2019) Health Statistics (2010) Kim (2018) al (2020)

Cerdá, M., Johnson-Lawrence, V. D., & Galea, S. (2011). Lifetime income patterns and alcohol consumption: investigating the association between long- and short-term income trajectories and drinking. Social science & medicine (1982), 73(8), 1178–1185. https://doi.org/10.1016/j.socscimed.2011.07.025

al, Matthew Ridley et. 2020. "Poverty, Depression, and Anxiety: Causal Evidence and Mechanism." https://www.science.org/doi/10.1126/science.aay0214.

Health Statistics, The National Center for. 2010. "The National Health and Nutrition Examination Surveys." https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2009.

Kim, & von dem Knesebeck, T. J. 2018. "Income and Obesity: What Is the Direction of the Relationship? A Systematic Review and Meta-Analysis." https://bmjopen.bmj.com/content/8/1/e019862.

Pruim, Randall. 2015. *National Center for Health Statistics: Data from the US National Health and Nutrition Examination Study*. https://www.cdc.gov/nchs/nhanes/index.htm?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fnchs%2Fnhanes.htm.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.