STA302 Fall 2023 Methods of Data Analysis 1
Final Project Proposal (Part 1)

Names of Group Members	Contribution to Proposal
Dian Rong	Model fitting
Krishna Kumar	Paper citations
Carter Ponce	Assessment of assumptions
Jing Yu	Preparation of Data

Dataset from Proposal pg: https://cran.r-project.org/web/packages/NHANES/NHANES.pdf
Notes on Research Question:

- https://www.canva.com/design/DAFvjygqh_g/aiSXFfquUi4O_X1uOpsKiQ/view

Sections:

1. Preparation of Data
2. Model fitting
3. Paper citations
4. assessment of assumptions

A. Research Question and Supporting Literature

1. *What is the research question you will be studying in this project? Be sure to explicitly refer to the variables under study and avoid using vague language to describe your study question.*

How do the following factors impact (if at all) an individual's ratio of family income to poverty (INDFMPIR), for individuals at least 30 years old, in the United States from 2017 to 2018?

- **BMI**
- **MaritalStatus**
- **How many hours the individual spends sleeping**
- **How much Alcohol the individual consumes everyday**
- **Whether the individual currently Smokes regularly**

2. *Provide an explanation for why a linear regression model would allow you to answer your research question. What aspect of your fitted model would give you the answer.*

A linear regression model allows us to answer our research question as it finds the relationship between BMI, marital status, sleep, alcohol, and smoking, on a person's ratio of income. Specifically, by taking the ratio of income to poverty guidelines, we are able to contextualize the income of the individual and their socioeconomic class. Then we can see, with all other factors held constant, how each factor affects the income to poverty ratio. This can also allow us to compare each factor and see which has the largest impact on expected income to poverty ratio.

The coefficients of each factor will give us the answer of how large of an impact each factor has on the expected income to poverty ratio: The higher the ratio, the higher the income. We should note that the ratio ranges from 1 to 5.

3. *Provide proper citations for 3 peer-reviewed academic research articles related to your specific research question or your topic of interest. For each, describe how the results of the article relate to your research question. Further, rank each article on a scale of 1 to 3 (1=not useful, 2=slightly useful, 3=very useful) based on how useful the article is in providing insight into the population relationship you wish to estimate. Justify this ranking.*

Citation	Description, ranking and justification
Ormond, Gillian, and Raegan Murphy. “The Effect of Alcohol Consumption on Household Income in Ireland.” Alcohol, vol. 56, Elsevier BV, Nov. 2016, pp. 39–49. https://doi.org/10.1016/j.alcohol.2016.10.003.	1. This study concludes that moderate and heavy drinkers have higher household incomes than non-drinkers and individuals who never drank.
Dunga, Steven Henry. “A GENDER AND MARITAL STATUS ANALYSIS OF HOUSEHOLD INCOME IN a LOW-INCOME TOWNSHIP.” Questa Soft, 2017, www.ceeol.com/search/article-detail?id=531140.	2. This study concludes that married households have the highest monthly income, compared to other households (widowed, divorced, single male/female, etc.). It also concludes that widowed females have the lowest monthly household income of all the groups.
Ogden et al.. “Prevalence of Obesity Among Adults, by Household Income and Education — United States, 2011–2014.” Morbidity and Mortality Weekly Report, vol. 66, no. 50, Centers for Disease Control and Prevention, Dec. 2017, pp. 1369–73. https://doi.org/10.15585/mmwr.mm6650a1..	3. This study concludes that among women, the prevalence of obesity decrease as income increased. Among men, the prevalence of obesity was lower in the lowest and highest income groups, and highest in the middle income group. The study concludes that obesity is more prevalent in lower income households, due to household income as well as additional variables such as physical activity facilities, availaibility of healthy/fresh food, and the general environment.

4. *Provide the database/library where you located the above academic papers. List the search terms used to find these papers, in addition to the number of results for each search term.*

Database/library searched	Search terms used	Number of results for each

Sciencedirect.com Scholar.google.com	- “(correlation between) BMI and household income” - “(relationship between) marital status and household income ” - “(How does) alcohol consumption affect household income”	- 162,000 - 2,840,000 - 445,000

B. Data Description, Justifications and Summary

1. *Provide the website from which your chosen data was obtained/downloaded.*

Website:**	https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017

*** If your data was obtained from a data repository (e.g. Kaggle, UCI Repository, etc.), please state how your research question differs from the original purpose of these data.*

2. *List the variables you have selected to be part of your preliminary model (minimum of 5 with at least one a categorical variable). Please give an understandable name to each variable rather than writing the name that appears in R.*

   For each variable, justify why you have chosen to use this variable over others in the dataset, and what the role of each variable will be (e.g., predictor of interest, predictor informed by literature, confounder, etc.).

Variable Name	Justification for Use	Role in Model
MaritalStatus	Marital status is an important socioeconomic variable. For example, married individuals may benefit from shared living expenses and potentially higher household income compared to other statuses.	predictor informed by literature
BMI	BMI is a health-related index. Individuals with different BMIs may experience different health issues, leading to potentially different employment opportunities and work productivity.	predictor informed by literature

SleepHrsNight	Sleep duration can impact an individual’s well-being and work productivity. Insufficient sleep may reduce productivity.	predictor of interest
AlcoholDay	The amount of alcohol consumption may affect an individual’s health and lifestyle choice. For example, excessive alcohol consumption may lead to health problems, absenteeism from work, and decreased job performance.	predictor informed by literature
SmokeNow	Smoking can have significant health and lifestyle implications, potentially affecting an individual’s employment opportunities.	predictor informed by literature

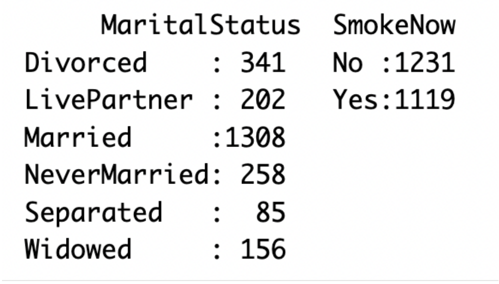
3. *Produce a table of numerical summaries of the variables listed above. Summaries should be appropriate to the type of variable, and interesting/important characteristics about variables should be mentioned in an informative caption. Include your summary table below.*

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

Table 1: This table provides summary statistics for numerical variables, including Mean, Median, SD, IQR, skewness and presence of outliers.

Categorical variables:



```
      MaritalStatus  SmokeNow 
 Divorced    : 341   No :1231  
 LivePartner : 202   Yes:1119  
 Married     :1308             
 NeverMarried: 258             
 Separated   :  85             
 Widowed     : 156             
```

Table 2: This table provides a summary for categorical variables.

C. Preliminary Model Results

1. *Fit your preliminary multiple linear model and present the estimated relationship. Present this information carefully so that it is easily readable and understandable.*

```
lm(formula = IncomeRatio ~ MaritalStatus + BMI + SleepHrsNight + 
    AlcoholDay + SmokeNow, data = new_data)

Coefficients:
              (Intercept)   MaritalStatusLivePartner       MaritalStatusMarried
                  3.70259                    0.02734                    0.91413
MaritalStatusNeverMarried     MaritalStatusSeparated       MaritalStatusWidowed
                 -0.02435                   -0.14739                    0.25004
                      BMI              SleepHrsNight                 AlcoholDay
                 -0.01336                   -0.08211                   -0.07836
              SmokeNowYes
                 -0.45728
```

In words, estimated relationship is IncomeRatio = 3.70259 + 0.02734 * (MaritalStatus is LivePartner)
+ 0.91413 * (MaritalStatus is Married) -0.02435 * (MartialStatus is NeverMarried)
-0.14739 * (MaritalStatus is Separated) + 0.25004 * (MartialStatus is Widowed)
-0.01336 * BMI -0.08211 * SleepHrsNight -0.07836 * (average number of drinks on a day the participant is drinking)
-0.45728 * (currently smoke)

2. *Justify your choice of how you included the categorical variable in your preliminary model. How does this choice contribute to answering your research question?*

MaritalStatus and SmokeNow are categorical variables.

They are all included as indicator variables. This is because, at this point, we are solely interested in how they independently affect IncomeRatio. For example, we want to know if being married, when all other factors are held constant, has a higher expected household income to poverty ratio, than being widowed. We want to similarly compare the other marital statuses, and those who currently smoke or don't.

3. *Do your estimated coefficients align/agree with the results of your three peer-reviewed articles? Explain in what way they differ/agree and provide a reason why this might be the case.*

The estimated coefficient of AlcoholDay (which measures the average number of alcoholic beverages consumed on a day when the individual drinks alcohol) is -0.07836. This contradicts Ormand, et al. (2016) which suggests that moderate and heavy drinkers have higher household incomes than non-drinkers and individuals who have never drank. This could be because our model doesn't differentiate between the effect of the 1st drink and the 4th drink.

The estimated coefficient of 0.91413 for the married marital status is larger than the coefficients for all the other marital status. This agrees with the result from Dunga (2017). This could be because the partnership and commitment needed for marriage may be correlated to higher household incomes.

The estimated coefficient of -0.01336 for BMI suggests higher BMIs are correlated with lower incomes which agrees with Ogden et al. (2017). This could be because those with higher BMIs may face discrimination due to their weight, causing decreased income and therefore a lower income to poverty ratio.

4. *Perform a complete assessment of the assumptions of your preliminary model. Do you observe violations of assumptions or conditions? Describe how you came to this conclusion, making explicit reference to any plots or other information that is relevant.*

Across all five of our predictor variables, we consistently validate the first three assumptions of linear regression:

1) Linearity of the relationship (mean zero) assumption:
    a) $E(\varepsilon|X) = 0$
2) Uncorrelated errors (independence) assumption:
    a) $Cov(\varepsilon_i, \varepsilon_j) = 0$
3) Constant error variance (homoscedasticity) assumption:
    a) $Var(\varepsilon|X) = \sigma^2 I$

However, also consistently, we find that our data violates the Normality Assumption of linear regression:

4) Normal errors assumption:
    a) $\varepsilon \mid X \sim N_n(0, \sigma^2 I)$

We can validate the **Linearity Assumption** since we find no signs of any other trends (curves, log functions, etc.)

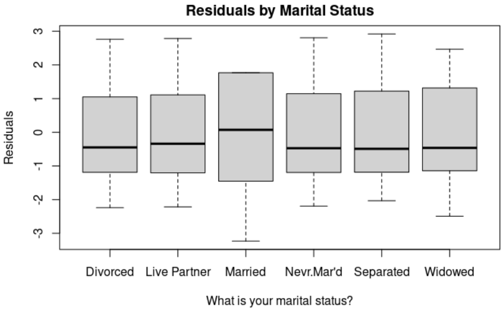
We can validate the **Uncorrelated Errors Assumption** since we find no significant grouping of residual points (disregarding the discrete nature of our variables).

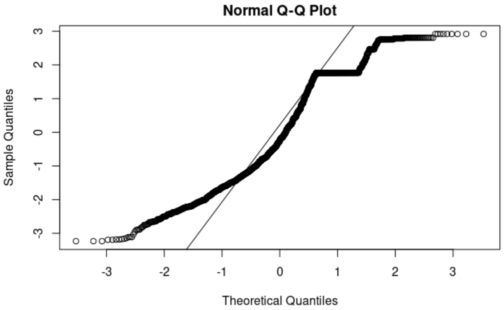
We can validate the **Constant Error Variance Assumption** since we cannot detect any signs of the *triangle pattern* of residual point scatter in our plots.

We can NOT validate the **Normal Errors Assumption** since there is *significant* deviation from the diagonal line in all of our QQ plots. This poses problems for our analysis as we may have concerns regarding biased parameter estimates, incorrect confidence intervals, and inaccurate hypothesis tests.

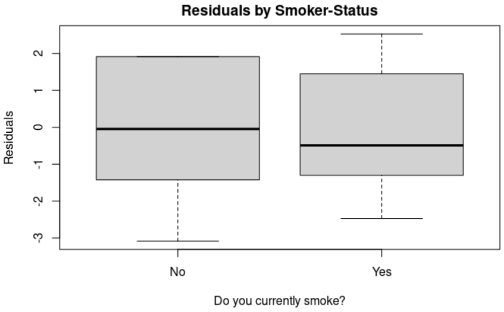
5. *Include all relevant plots created for assessing model assumptions below, with appropriate axis labels and captions.*

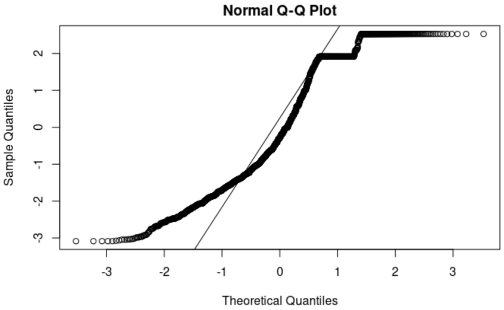
Marital Status



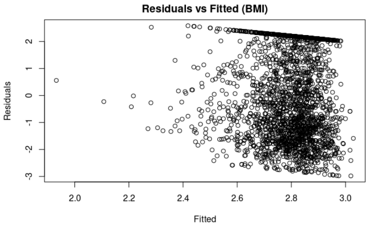


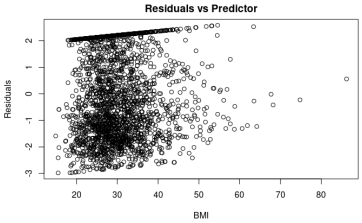
Smoker Status

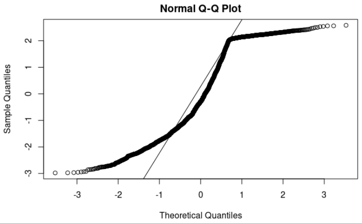




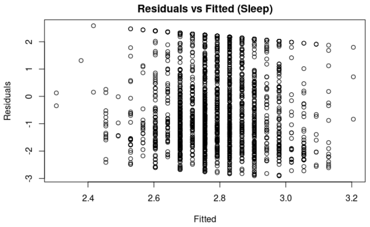
BMI

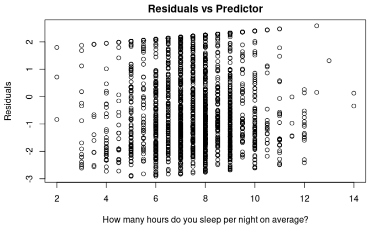


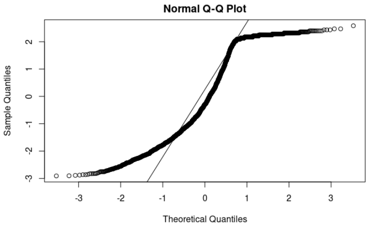




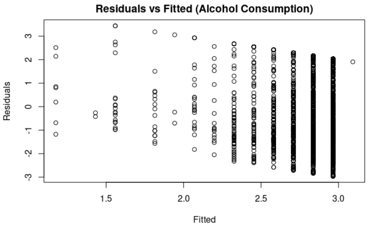
Hours of Sleep Per Night

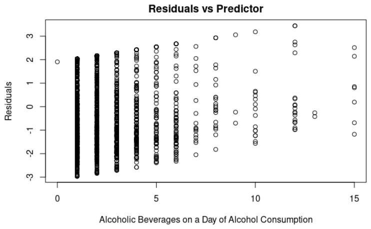


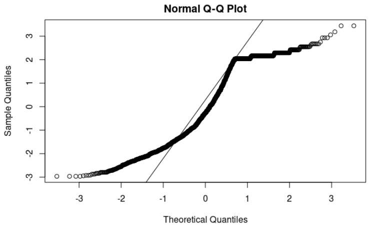




Alcohol Consumption


Residuals vs Fitted (Alcohol Consumption)
Residuals
3
2
1
0
-1
-2
-3
1.5
2.0
2.5
3.0
Fitted


Residuals vs Predictor
Residuals
3
2
1
0
-1
-2
-3
0
5
10
15
Alcoholic Beverages on a Day of Alcohol Consumption


Normal Q-Q Plot
Sample Quantiles
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
Theoretical Quantiles