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

Names of Group Members	Contribution to Proposal
Jiahao Gu	Finding the dataset, 1A, 1B, and part B
Zilong Zhao	Finding the first two research papers, A3 and A4
Pengfeng Yang	Finding the last research paper, A3, A4 and C1
Fanke Qin	Complete C2,C3
Hilary Zou	Complete C4, C5

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 daily habits such as physical activity, smoking status, alcohol consuming, sleeping status, and drug usage impact health conditions, measured by self-reported number of days participant's physical health was not good?

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.*

Linear regression allows us to study the impact of multiple independent variables (multiple aspects of daily habits) on a single dependent variable (number of days with poor physical health). This makes it an ideal tool to analyze how each habit contributes to health outcomes. The parameter vector inherent in our fitted model serves is important for explaining the relationship between predictors and the response variable. Each coefficient within this vector quantifies the significance of the associated predictor, thereby providing insights into their individual contributions to the dependent variable.

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.*

Citation	Description and ranking
Humphreys, B. R., McLeod, L., & Ruseski, J. E. (2014). PHYSICAL ACTIVITY AND HEALTH OUTCOMES:	The article finds a significant association between physical activity and health conditions. Any level of regular physical activity, compared to inactivity,

EVIDENCE FROM CANADA. Health Economics, 23(1), 33–54. https://doi.org/10.1002/hec.2900	reduces the probability of having diabetes, high blood pressure, arthritis, and reporting being in fair or poor health. This aligns with the focus of our research question, which seeks to understand how daily habits, such as physical activity, affect physical health. **Rank 3**
Babizhayev, M. A., & Yegorov, Y. E. (2011). Smoking and health: association between telomere length and factors impacting on human disease, quality of life and life span in a large population-based cohort under the effect of smoking duration. Fundamental & Clinical Pharmacology, 25(4), 425–442. https://doi.org/10.1111/j.1472-8206.2010.00866.x	The article explicitly associates smoking with a multitude of serious health conditions such as cancer, coronary artery disease, and autoimmune disorders, aligning with our research question to study the relationship between daily habits like smoking and health. **Rank 3**
O'Keefe, J. H., Bybee, K. A., & Lavie, C. J. (2007). Alcohol and Cardiovascular Health. The Razor-Sharp Double-Edged Sword. Journal of the American College of Cardiology, 50(11), 1009–1014. https://doi.org/10.1016/j.jacc.2007.04.089	The article presents evidence of J-shaped associations between alcohol intake and various health outcomes. This means that low to moderate consumption might be associated with reduced risks or benefits, but as consumption increases, the risks also increase. This relates to our research question because it also studies the relationship between alcohol consumption and health. **Rank 3**

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
University of Toronto Libraries	physical activities and health	519,725
University of Toronto Libraries	smoking and health	303,765
University of Toronto Libraries	alcohol and health	320,322

B. Data Description, Justifications and Summary

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

Website:**	https://cran.r-project.org/web/packages/NHANES/NHANES.pdf

*** 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 <u>each</u> 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
Self-reported Number of Days Participant's Physical Health was NOT Good Out of the Past 30 Days (DaysPhysHlthBad)	The self-reported number of days a participant's physical health was not good out of the past 30 days reflects physical health conditions because it provides a subjective measure of an individual's perceived health status, revealing the frequency of days impacted by poor health, which can help in identifying physical health conditions.	Continuous Response of Interest
Average Frequency of Physical Activity (PhysActiveDays)	Engaging in physical activities enhances physical health conditions as it promotes cardiovascular fitness and improves immune function, thereby mitigating the risk of chronic diseases and fostering overall body resilience. We think that people who do physical activities frequently are likely to be in better physical health conditions.	Continuous Predictor of Interest
Average Amount of Alcohol Consumed (AlcoholDay)	Consuming alcohol affects physical health conditions as it can impair liver function, weaken the immune system, increase the risk of chronic diseases such as cardiovascular disease and liver cirrhosis. We think that people who consume more alchohol are likely to be in worse physical health conditions.	Continuous Predictor of Interest
Average Sleeping Hours (SleepHrsNight)	Adequate sleep is crucial for physical health as it facilitates essential bodily functions such as cellular repair, which reduces the risk of chronic diseases. We think that people who sleep more are	Continuous Predictor of Interest

	likely to be in better physical health conditions.	
Drug Usage (Smoke100n)	Using drugs can adversely affect physical health conditions because they can alter bodily functions, lead to addiction and cause a range of harmful short- and long-term health consequences. We think that people who use drugs are likely to be in worse physical health conditions.	Categorical Predictor of Interest
Smoking Status (HardDrugs)	Smoking negatively affects physical health conditions as it introduces a multitude of harmful chemicals into the body, leading to the development of various diseases including cancer and cardiovascular disease. We think that people who smoke are likely to be in worse physical health conditions.	Categorical Predictor of Interest

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.*

Variable Name in the Dataset	**Minimum**	**Median**	**Mean**	**Maximum**	**Interesting/Important Characteristics**
DaysPhysHlthBad	0.000	0.000	3.366	30.000	Based on the histogram of DaysPhysHlthBad, the response shows a strong pattern of right-skewness.
PhysActiveDays	0.000	0.000	1.738	7.000	Base on the scatter plot "DaysPhysHlthBad vs. PhysActiveDays", there appears to be a negative relationship between DaysPhysHlthBad and AlcoholDay.
AlcoholDay	1.000	2.000	3.401	82.000	Base on the scatter plot "DaysPhysHlthBad vs. AlcoholDay", there appears to be no relationship between DaysPhysHlthBad and AlcoholDay. There is an outlier with a value of 82, which

					deviates significantly from the other data.
SleepHrsNight	2.000	7.000	6.791	12.000	Base on the scatter plot "DaysPhysHlthBad vs. SleepHrsNight", there appears to be a negative relationship between DaysPhysHlthBad and SleepHrsNight.
Variable Name in the Dataset	**Number of "1"s**		**Number of "0"s**		**Interesting/Important Characteristics**
Smoke100n	Smokers: 564		Non-Smokers: 621		Base on the box plot "DaysPhysHlthBad vs. Smoke100", smokers tend to have higher DaysPhysHlthBad.
HardDrugs	Yes: 245		No: 940		Base on the box plot "DaysPhysHlthBad vs. HardDrugs", drug users tend to have higher DaysPhysHlthBad.

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.*

$y_i = 4.3040 - 0.1126x_1 - 0.0335x_2 + 0.0184x_3 + 0.7832x_4 + 0.2557x_5 - 2.5936x_6 - 0.1090x_7 + 3.5447x_8 - 0.6727x_9$

Variables:

y_i: DaysPhysHlthBad;

x_1: PhysActiveDays, x_2: AlcoholDay, x_3: SleepHrsNight, x_4: Smoke100n (1 for smokers and 0 for non-smokers), x_5: HardDrugs (1 for yes and 0 for no), x_6: PhysActive (1 for yes and 0 for no), x_7: Marijuana (1 for yes and 0 for no), x_8: SleepTrouble (1 for yes and 0 for no), x_9 : Alcohol12PlusYr (1 for yes and 0 for no).

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?*

By using drug usage, smoking status, physical activeness, and drinking status as categorical variables, we can conduct statistical analysis to quantify the relationship between them and physical health status. This indicates the likelihood of drug users, smokers, inactive

individuals, or alcohol consumers having poorer physical health compared to non-drug users, non-smokers, physically active individuals, or those who don't consume alcohol. By evaluating their importance and contribution, we can understand the self-reported differences in physical health status. This provides evidence on whether drug use, smoking status, physical inactivity, and alcohol consumption are related to poor physical health. By including these categorical variables, we can construct a more comprehensive model that considers multiple potential factors affecting physical health. This helps us better understand the impact of these lifestyle factors on health.

3. *What is the contribution of your research question. For example, is the goal of your project to have a high prediction accuracy? Is the goal of your project to draw strong conclusions about the overall population? Briefly explain your plan.*

Our project goal is drawing strong conclusions about the overall population and we hope to make general inferences about a larger population based on our research. After fitting the linear regression model, we know that, if the coefficient of a variable is positive, then it is likely that daily habit will positively affect physical health, otherwise, it will negatively affect physical health. How large the coefficient is also tells us about how significant that habit is to physical health.

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.*

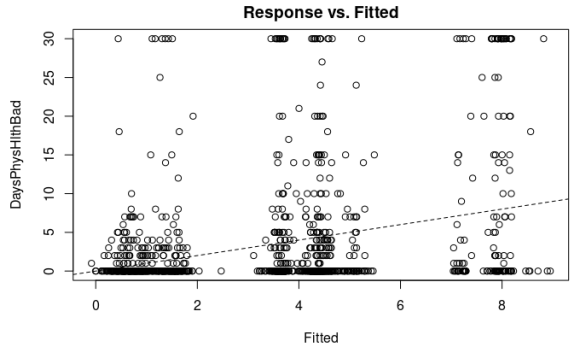
The Response vs. Fitted pot does not show a clear linear trend. There are plenty of outliers away from the fitted line in the graph. Then, as of the graph of pairwise scatterplot, there are no obvious pattern except linear patterns. As of result based on the conditions from the Response vs. Fitted graph, multiple linear regression conditions are not satisfied, so the result of residuals plot may be unreliable.

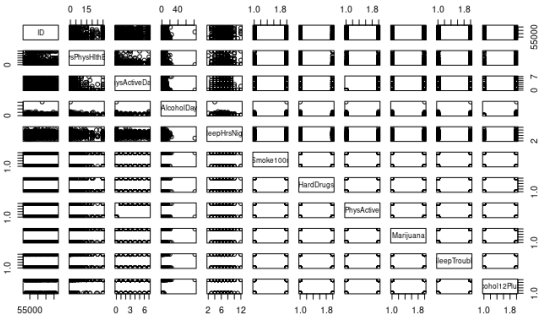
In the plots of Residuals vs each predictor and residual vs fitted, there is no obvious systematic pattern, including curves, other functions of predictors, and fanning pattern, or large clusters of many points, indicating that there is no strong evidence supporting violations to Linearity, Uncorrelated Errors and Constant Variance.

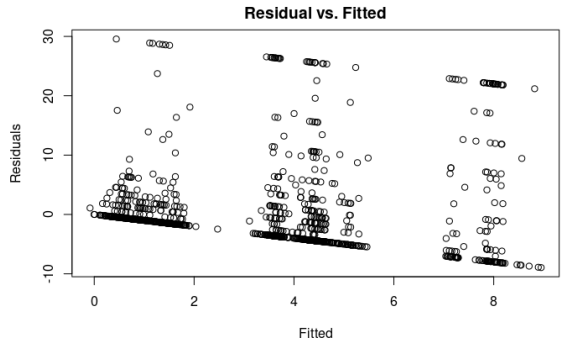
Using the Normal Q-Q Plot, the data are clearly not normally distributed. Even though plenty of data are lying on top of the fitted line in the diagram, there are also huge amount of data outside the fitted line. There are excessive amounts of outliers that indicates the data are not normally distributed.

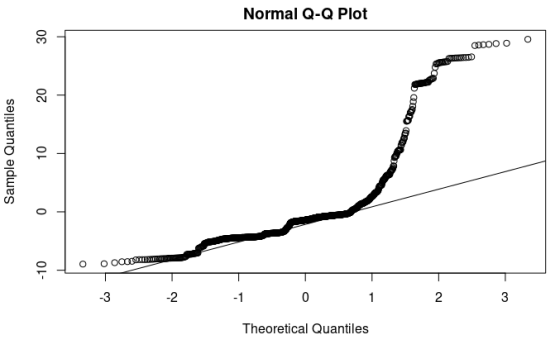
To summarize the observations from the plots, multiple linear regression conditions are not satisfied and there are strong evidence of violations to Normality. As a result of the violations, our estimators will not have a Normal sampling distribution.

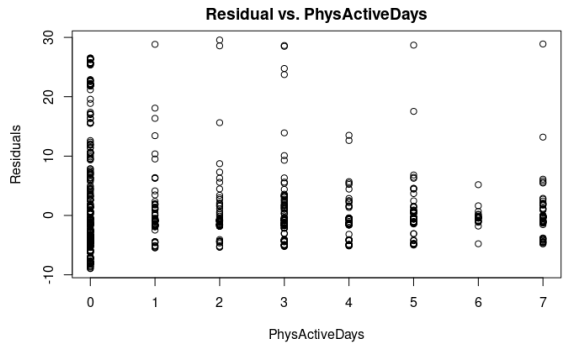
5. *Include all relevant plots created for assessing model assumptions below, with appropriate axis labels and caption.*

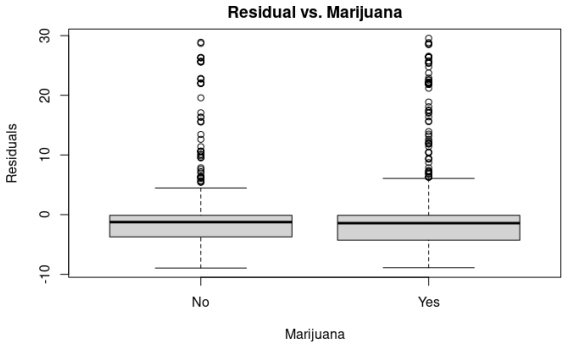

Response vs. Fitted
DaysPhysHlthBad
Fitted

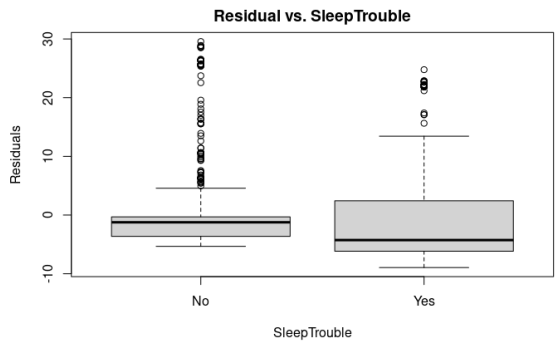

ID
DaysPhysHlthBad
DaysActiveDays
AlcoholDay
SleepHrsNight
Smoke100
HardDrugs
PhysActive
Marijuana
SleepTrouble
Alcohol12PlusYr


Residual vs. Fitted
Residuals
Fitted

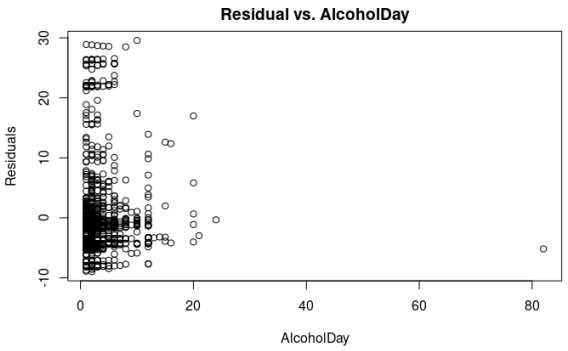

Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles

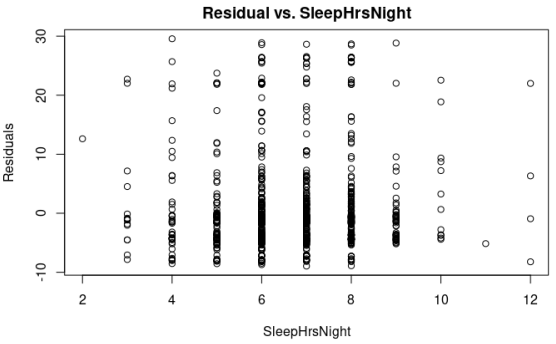

Residual vs. PhysActiveDays
Residuals
PhysActiveDays

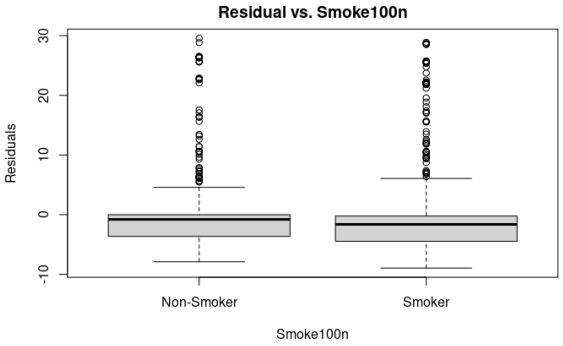

Residual vs. Marijuana
Residuals
No
Yes
Marijuana

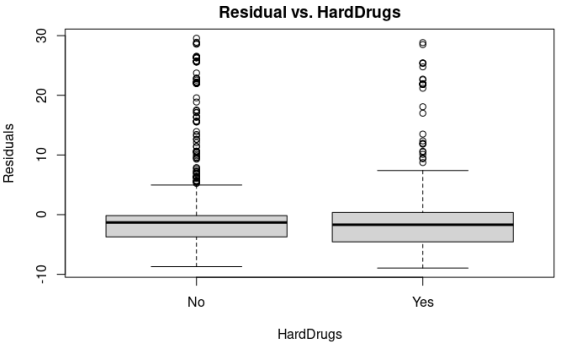

Residual vs. SleepTrouble
Residuals
No
Yes
SleepTrouble

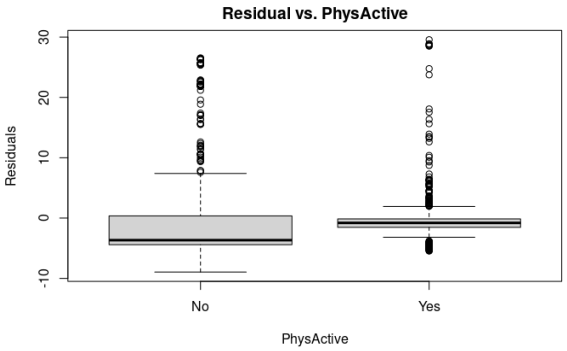

Residual vs. Alcohol12PlusYr
Residuals
No
Yes
Alcohol12PlusYr


Residual vs. AlcoholDay
Residuals
AlcoholDay
-10
0
10
20
30
0
20
40
60
80


Residual vs. SleepHrsNight
Residuals
SleepHrsNight
-10
0
10
20
30
2
4
6
8
10
12


Residual vs. Smoke100n
Residuals
Smoke100n
-10
0
10
20
30
Non-Smoker
Smoker


Residual vs. HardDrugs
Residuals
HardDrugs
-10
0
10
20
30
No
Yes


Residual vs. PhysActive
Residuals
PhysActive
-10
0
10
20
30
No
Yes