BIST0610 Project Proposal

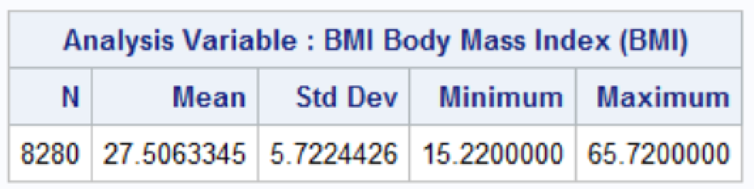
Kevin Zheng, Phoebe Chu, Marisha Shanthikumar
March 8, 2022

Introduction

The purpose of this study is to examine how certain factors affect body mass index in adults over the age of 18 and under the age of 85. The National Center for Health Statistics (NCHS) conducts the National Health Interview Survey (NHIS) each year to monitor the health of the United States population. Investigators conduct face-to-face household interviews and follow-up with telephone interviews, if necessary. Telephone interviews were also conducted if requested by the respondent or if there was difficulty traveling to the participant's home. The data used in this study is derived from the 2018 Sample Adult questionnaire, in which one adult per family is randomly selected. The sample adult responds for themselves unless they are physically or mentally unable to do so, in which a knowledgeable proxy is allowed to answer for the adult.

About the Data

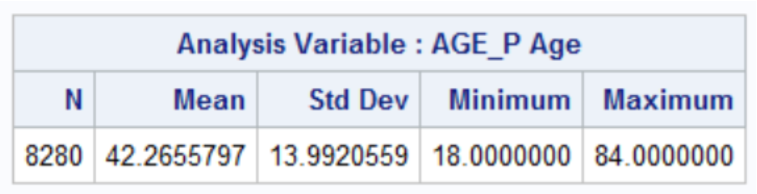
Our continuous response variable is body mass index (BMI), a person's weight in kilograms divided by the square of height in meters. It holds information on a person's weight category. A summary of the data set sample after removing samples with unknown values is given below.



Analysis Variable : BMI Body Mass Index (BMI)				
N	Mean	Std Dev	Minimum	Maximum
8280	27.5063345	5.7224426	15.2200000	65.7200000

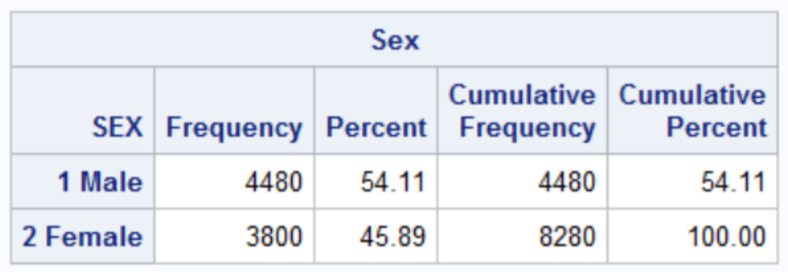
We have chosen seven covariates to examine in this study. Below is the description and summary for each of these variables, after missing values were removed from the dataset. We consider Refused, Not Ascertained, and Don't Know as missing values.

- Age (AGE_P) - continuous from age 18 to 84; we have excluded those 85 and above because their specific age is not available in this dataset



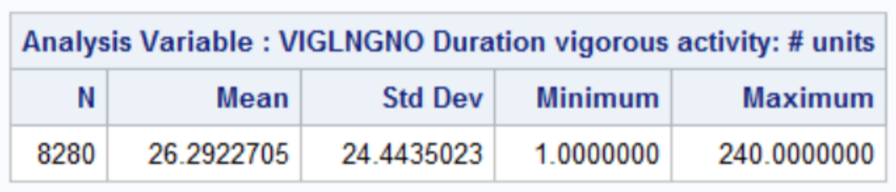
Analysis Variable : AGE_P Age				
N	Mean	Std Dev	Minimum	Maximum
8280	42.2655797	13.9920559	18.0000000	84.0000000

- Sex (SEX) - possible values are Male or Female



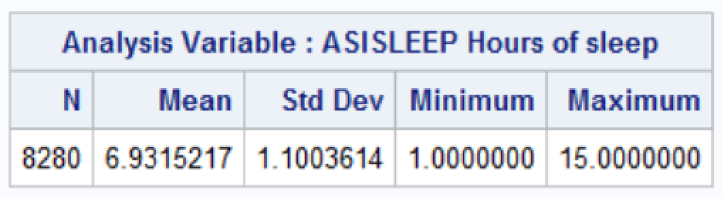
Sex				
SEX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 Male	4480	54.11	4480	54.11
2 Female	3800	45.89	8280	100.00

- Duration of vigorous activity (VIGLNGNO) - the average duration in minutes that each person performs vigorous activity each time they exercise



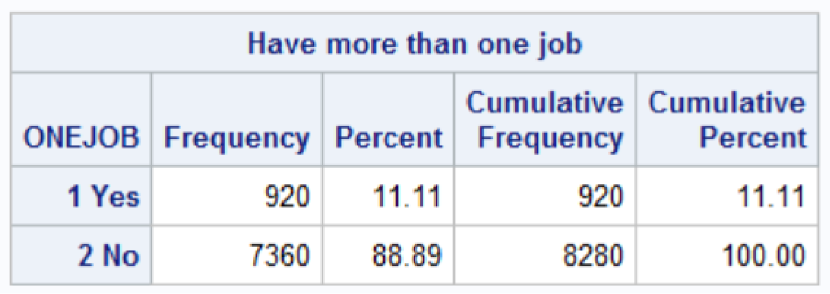
Analysis Variable : VIGLNGNO Duration vigorous activity: # units				
N	Mean	Std Dev	Minimum	Maximum
8280	26.2922705	24.4435023	1.0000000	240.0000000

- Hours of sleep (ASISLEEP) - how many hours of sleep on average in a 24-hour period



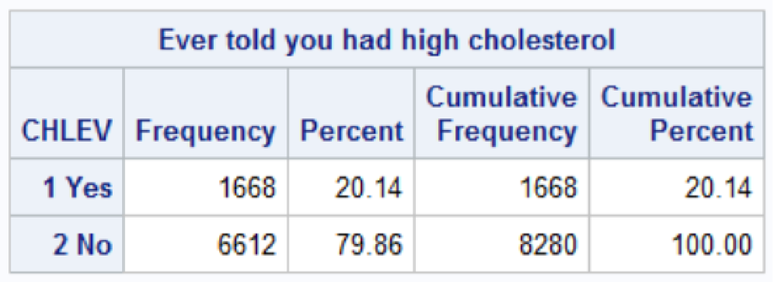
Analysis Variable : ASISLEEP Hours of sleep				
N	Mean	Std Dev	Minimum	Maximum
8280	6.9315217	1.1003614	1.0000000	15.0000000

- Have more than one job (ONEJOB) - whether or not the person has more than one job with values yes or no



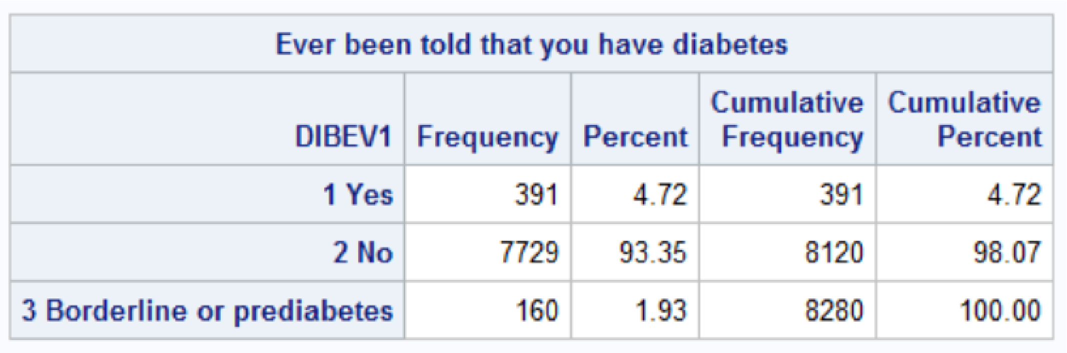
Have more than one job				
ONEJOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 Yes	920	11.11	920	11.11
2 No	7360	88.89	8280	100.00

- Ever had high cholesterol (CHLEV) - whether or not the person has ever been diagnosed with high cholesterol with values yes or no



Ever told you had high cholesterol				
CHLEV	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 Yes	1668	20.14	1668	20.14
2 No	6612	79.86	8280	100.00

- Ever had diabetes (DIBEV1) - whether or not the person has ever been diagnosed with diabetes with values yes, no, or borderline/prediabetes



Ever been told that you have diabetes				
DIBEV1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1 Yes	391	4.72	391	4.72
2 No	7729	93.35	8120	98.07
3 Borderline or prediabetes	160	1.93	8280	100.00

Our final sample size is 8280.

Hypotheses

We have included a hypothesis for each of our covariates.

Age:

H_0: BMI is not significantly associated with age vs. H_A: BMI is significantly associated with age.

Sex:

H_0: BMI is not significantly different between sex vs. H_A: BMI is significantly different between sex.

Duration Physical activity:

H_0: BMI is not significantly associated with a respondent's duration of physical activity when they exercise vs. H_A: BMI is significantly associated with a respondent's duration of physical activity when they exercise.

Hours of sleep:

H_0: BMI is not significantly associated with a respondent's hours of sleep in 1 day vs. H_A: BMI is significantly associated with a respondent's hours of sleep in 1 day.

Having more than one job:
H_0: Whether or not someone works more than one job is not associated with a significant difference in BMI vs H_A: Whether or not someone works more than one job is associated with a significant difference in BMI.

Having high cholesterol:
H_0: Whether or not someone has been told they have high cholesterol is not associated with a significant difference in BMI vs H_A: Whether or not someone has been told they have high cholesterol is associated with a significant difference in BMI.

Having Diabetes:
H_0: Whether or not someone has been told they have diabetes is not associated with a significant difference in BMI vs H_A: Whether or not someone has been told they have diabetes is associated with a significant difference in BMI.

Statistical Methods

We plan to conduct the following analyses:

- Exploratory Data Analysis - we will carefully look at each of our variables to discover any patterns or anomalies. For continuous covariates of age, duration of physical activity, and hours of sleep, scatterplots will be made comparing each of these covariates with BMI. We will also examine the scatterplots between these covariates to see if they have any relationship. For categorical variables of sex, having more than one job, ever having high cholesterol, and ever having diabetes, boxplots will be made for each of these variables, which will compare the BMI distribution for each variable's factors.

- Model diagnostics - we will check assumptions on the model and make necessary adjustments, if possible
  - Jackknife and/or studentized residual plots for influential points, violation of constant variance
  - For influential points, we may look at Cook's distance, DFBETAS, DFFITS, and/or COVRATIOS
  - Check normality assumption where possible

- Hypothesis tests - we will conduct hypothesis testing on our covariates to determine if they make significant contributions to the model. For each covariate, the null hypothesis is that the coefficient β_i=0, and the alternative hypothesis will be that $\beta_i \neq 0$. Then, the significance of these coefficient estimates will be tested for each covariate, and we will observe the corresponding p-values.

- Model Building - We will try to run different models on our dataset
  - Run a multiple linear regression (MLR) with all of the covariates
  - Check slope estimates and corresponding p-values for each variable
  - Check ANOVA and F test for all of the covariates

- Regression modeling -
  We will examine linear regression, multiple regression, and non-linear models, depending on the necessity, to describe the relationships between our response variable and the set of independent variables. We intend to mainly use multiple linear regression while including all the covariates simultaneously, and make adjustments as necessary.

  We might use stepwise regression modeling, since various input variables affect one output variable. We will build a model through forward stepwise regression by adding and inputting one variable at a time based on their significance and how it affects the target variable, where the addition of a variable should refine our model. We may also try backwards stepwise regression.

Sources
Survey Description:
https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2018/srvydesc.pdf

Data: https://www.cdc.gov/nchs/nhis/nhis_2018_data_release.htm

SAS Code to obtain dataset:

```
data raw;
	Set NHIS.Samadult (keep= AGE_P SEX VIGLNGNO ASISLEEP ONEJOB CHLEV DIBEV1 BMI);
	If AGE_P > 84 then delete; *Remove those age 85+;
	If VIGLNGNO =. then delete; *Remove missing values;
	If VIGLNGNO >= 998 then delete; *Remove Not Ascertained, Don't Know;
	If ASISLEEP >= 97 then delete; *Remove Refused, Not Ascertained, Don't Know;
	If BMI >= 99.99 then delete; *Remove Unknown;
	If ONEJOB =. then delete; *Remove missing values;
	If ONEJOB >= 7 then delete; *Remove Refused, Don't Know;
	If CHLEV >= 7 then delete; *Remove Refused, Don't Know;
	If DIBEV1 >= 7 then delete; *Remove Refused, Don't Know;
run;

proc contents data=raw;
run;

proc freq data=raw;
	tables SEX ONEJOB CHLEV DIBEV1;
run;
```

```
proc means data=raw;
        var AGE_P VIGLNGNO ASISLEEP BMI;
run;
```