STAT 528
Winter Quarter 2021
Dr. Erosheva
January 6, 2021

STAT 528
APPLIED STATISTICS CAPSTONE

Data from the Cardiovascular Health Study (CHS)

Data Confidentiality Agreement

Before downloading the data for Homework 1, you must download, electronically sign, and upload this form to Canvas. We will not make the data available until everyone in the class has signed.

I, _____Dongyang Wang_____, understand that the data we will analyze in the CHS-related project are confidential data on human subjects and should not be made public. I promise to keep the data confidential, sharing it with no one other than the instructor and the members of this class.

I further promise that I will delete all copies of these data within one week of the last day of classes this quarter, March 12, 2021.

Signed, this 12th day of January 2021

_____*Dongyang Wang*_____ (Signature).

Dongyang Wang
Professor Taeb
Stat 528
1/16/2023

HW1

Q1

The scientific question for the problem is whether exercise is associated with mortality of individuals aged 65 years and older, and more specifically due to cardiovascular risks. In the papers, the scientific questions are slightly different but similar to a certain extent. For the Fried paper, "The main objective of the study is to identify factors related to the onset and course of coronary heart disease and stroke. The Cardiovascular Health Study (CHS) is designed to determine the importance of conventional cardiovascular disease (CVD) risk factors in older adults, and to identify new risk factors in this age group, especially those that may be protective and modifiable." For the Siscovick paper, "The authors assessed the cross-sectional association between intensity of exercise in later life and coronary heart disease risk factors and subclinical disease among 2,274 men and women, 65 years of age and older, who were participants in the Cardiovascular Health Study (CHS) during 1989-1990."

The population is U.S. citizens aged 65 years and older, particularly those who are on the Medicare eligibility lists and meet the eligibility criteria, namely non-institutionalized and able to give informed consent. For convenience of sampling, the sample also expected participants to remain in the area for 3 years and excluded the wheelchair-bound and hospice/cancer treatment-undergoing individuals.

The response (outcome) is mortality, or if the people are still alive by the end of the study. Specifically, we might be interested in the cardiovascular risks regarding mortality.

The specific aims are to perform analysis on whether the baseline variables (characteristics of an individual) and exercise variables (exercise behaviors of an individual) will affect a person's mortality risks, especially regarding cardiovascular risks.

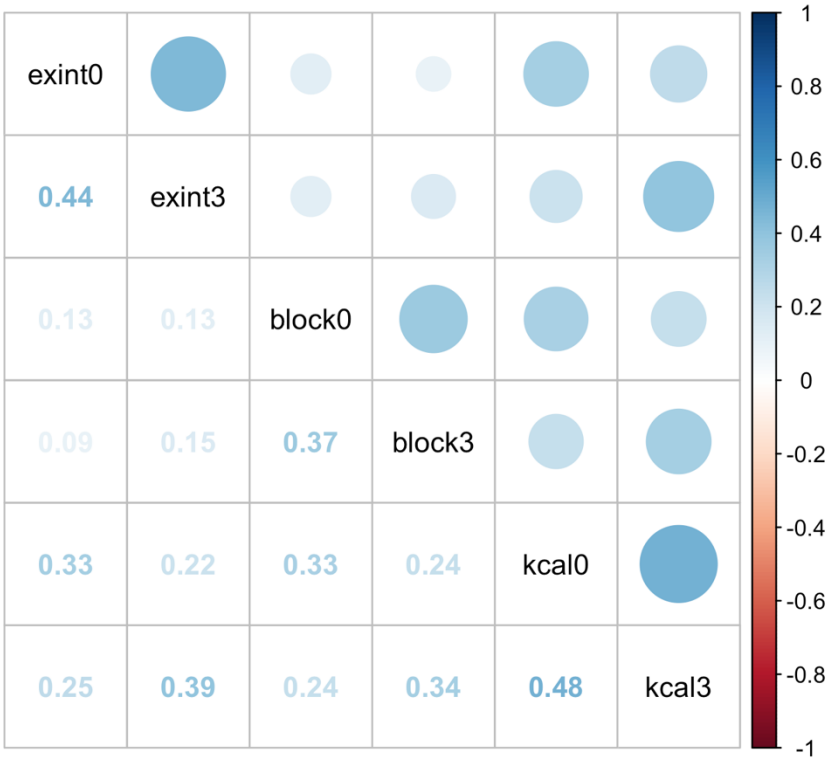
Q2

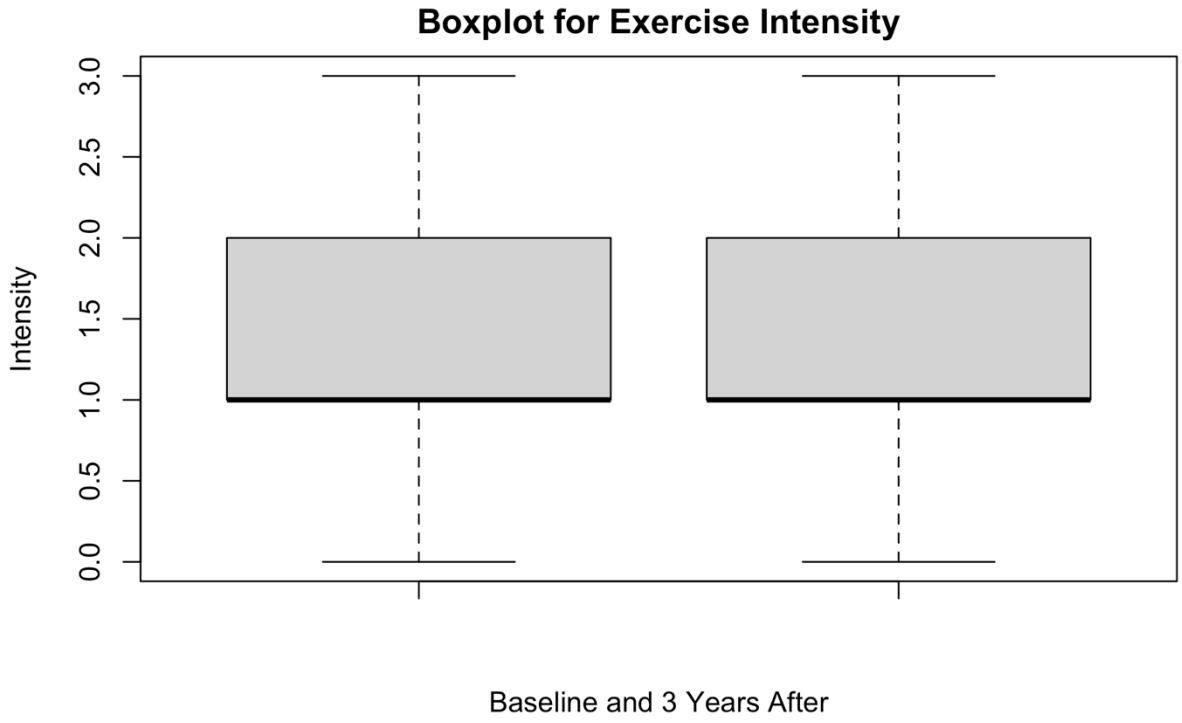
Part 1

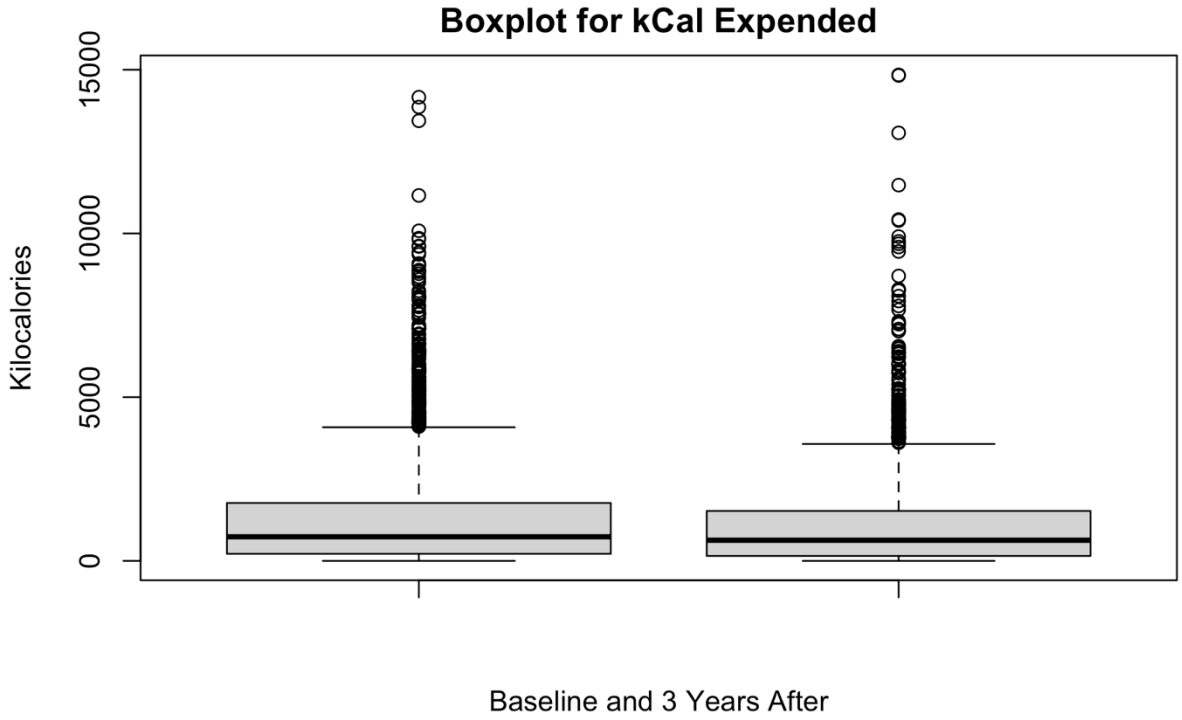
First, I cleaned up the dataset so the exercise variables can get rid of missing values. Because it makes no sense to consider the case if a person has died. Also, there are a lot of missing values due to unknown reasons, which would hinder the graphics and analysis.

The exercise variables are related to each other in terms of the following figures. The first graph shows that there are indeed correlations between the same category of exercise variables, e.g., exercise intensity. The correlation among different exercise variables is a bit lower, but the correlation among variables of the same year is still slightly higher. The three following boxplots

show that the exercise trends in general do not change much, but as people get older, the amount of calories expended does decreases. Furthermore, for each individual, I have calculated their difference of each category of exercise variables and outputted a table. People on average do not change level of exercise intensity but do walk more blocks, and less calories are expended.





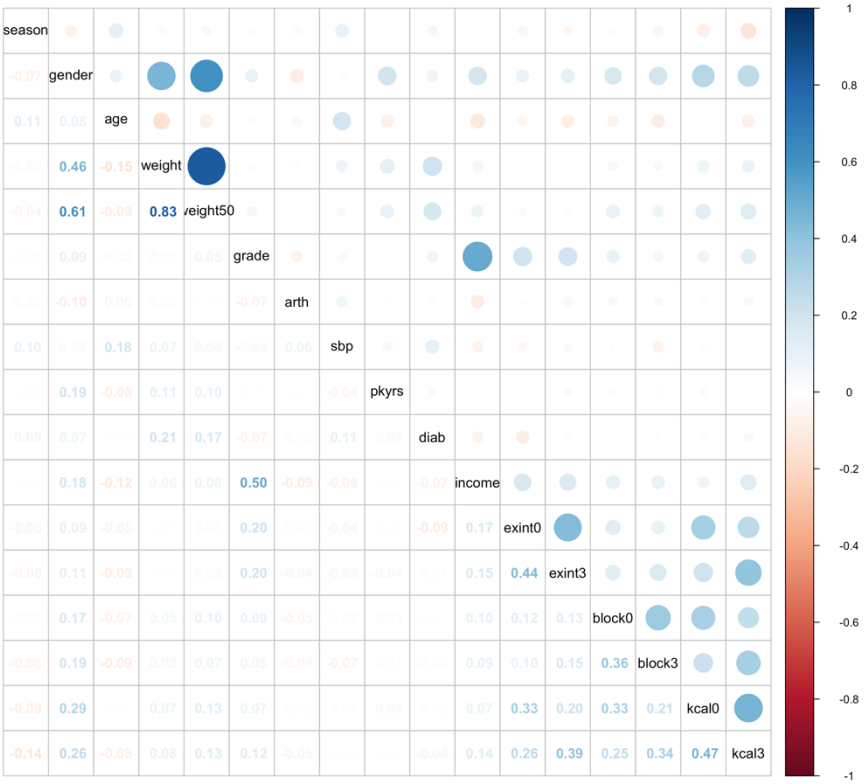


	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Exercise Intensity	-3.0	0	0.0	-0.06165971	0.000	3	283
Blocks Walked	-288.0	-24	0.0	5.22394221	24.000	696	502
kCal Expended	-12262.5	-675	-67.5	-211.58658763	314.875	14531	289

Part 2

First, I want to clarify the baseline variables. They include "season", "gender", "age", "weight", "weight50", "grade", "arth", "sbp", "pkyrs", "diab", "income". And the exercise variables are the same as in Part 1, namely "exint0", "exint3", "block0", "block3", "kcal0", "kcal3".

The following correlation plot shows that among the exercise variables and baseline variables (that have non missing values), the correlation is quite small. Gender seems the only baseline variable that has some slight association (above 0.2) with the calories expended but not other exercise activities. Moreover, grade and income are moderately correlated; weight and gender are moderately correlated; pervious weight and current weight are highly correlated.



Part 3

A simple linear regression returns us an output. This model uses mortality at the end of study as the dependent variable and a list of baseline variables as the independent variables. The result of the regression analysis is as follows. Significant variables include age, weight, weight at 50 years of age, smoking history, diabetes, and income (at 0.05 level, others all at 0.001 level). Based on the results, there seem a lot of factors unrelated to exercises but crucial to a person's healthy and mortality.

MODEL INFO:
Observations: 2086 (354 missing obs. deleted)
Dependent Variable: mortality
Type: OLS linear regression

MODEL FIT:
$F(11,2074) = 29.25$, $p = 0.00$
$R^2 = 0.13$
Adj. $R^2 = 0.13$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	-1.44	0.15	-9.80	0.00
season	0.00	0.01	0.03	0.97
gender	0.02	0.02	0.84	0.40
age	0.02	0.00	11.65	0.00
weight	-0.00	0.00	-4.12	0.00
weight50	0.00	0.00	4.13	0.00
grade	0.00	0.00	0.21	0.84
arth	-0.02	0.02	-1.50	0.13
sbp	0.00	0.00	1.65	0.10
pkyrs	0.00	0.00	7.41	0.00
diab	0.06	0.01	5.09	0.00
income	-0.01	0.01	-2.56	0.01

Part 4

After introducing exercise variables in the regression, only exercise intensity at 3 years of study seems to be significant at the 0.05 level among all exercise variables. But also note that there are fewer observations in this model since there are more missing values in the exercise variables.

With this being said, there still seems a correlation between exercise and mortality, since one level increase in the 3 year measure of exercise intensity is associated with a 0.03151 decreased chance in mortality. However, we cannot conclude for now with a causal relationship. There might be omitted variables that affect exercise and mortality at the same time. It may also be the case that healthy people tend to exercise more, so selection bias might be in place. As in the Siscovick paper, their conclusion is "The authors conclude that intensity of exercise in later life is associated with favorable coronary disease risk factor levels and a reduced prevalence of several markers of subclinical disease." So we know that association is in place but not necessarily causation.

Note: next page is the regression results for this part, and after it would be an appendix of the RMarkdown code I have written for this report.

References:

Fried, L.P. et al. (1991). The Cardiovascular Health Study: Design and Rationale. *Annals of Epidemiology*, **1**, 263–276.

Siscovick, D.S. et al. (1997). Exercise intensity and subclinical cardiovascular disease in the elderly. *American Journal of Epidemiology*, **145**, 977–986.

MODEL INFO:
Observations: 1668 (772 missing obs. deleted)
Dependent Variable: mortality
Type: OLS linear regression

MODEL FIT:
$F(17,1650) = 9.95$, $p = 0.00$
$R^2 = 0.09$
Adj. $R^2 = 0.08$

Standard errors: OLS

	Est.	S.E.	t val.	p
(Intercept)	-1.02	0.17	-6.15	0.00
season	0.01	0.01	0.81	0.42
gender	-0.00	0.02	-0.08	0.94
age	0.01	0.00	7.21	0.00
weight	-0.00	0.00	-3.23	0.00
weight50	0.00	0.00	3.71	0.00
grade	0.00	0.00	1.10	0.27
arth	-0.02	0.02	-0.99	0.32
sbp	0.00	0.00	0.67	0.51
pkyrs	0.00	0.00	6.22	0.00
diab	0.05	0.01	3.90	0.00
income	-0.01	0.01	-2.13	0.03
exint0	0.01	0.01	0.80	0.42
exint3	-0.03	0.01	-2.48	0.01
block0	-0.00	0.00	-0.21	0.83
block3	-0.00	0.00	-1.68	0.09
kcal0	0.00	0.00	1.22	0.22
kcal3	0.00	0.00	0.04	0.96

Stat 528 HW1

Dongyang Wang

2023-01-15

Question 1

```
rm(list=ls())
df = read.csv("CHSdataEx1.csv")
```

Question 2

Question 2.1

```
library(corrplot)
library(xtable)

df_exer = na.omit(df[,c("exint0", "exint3", "block0", "block3", "kcal0", "kcal3")])
corrplot.mixed(cor(df_exer),
               lower = "number",
               upper = "circle",
               tl.col = "black")

boxplot(df$exint0, df$exint3, main = "Boxplot for Exercise Intensity",
        xlab = "Baseline and 3 Years After",
        ylab = "Intensity")
tab1 <- summary(df$exint3- df$exint0)

boxplot(df$block0, df$block3, main = "Boxplot for Blocks Walked",
        xlab = "Baseline and 3 Years After",
        ylab = "Blocks")
tab2 <- summary(df$block3- df$block0)

boxplot(df$kcal0, df$kcal3, main = "Boxplot for kCal Expended",
        xlab = "Baseline and 3 Years After",
        ylab = "Kilocalories")
tab3 <- summary(df$kcal3- df$kcal0)

table1 <- rbind(tab1, tab2, tab3)
rownames(table1) <- c("Exercise Intensity", "Blocks Walked", "kCal Expended")
table1
```

Question 2.2

```
baseline = c("season", "gender", "age", "weight", "weight50", "grade", "arth", "sbp", "pkyrs", "diab",
exercise = c("exint0", "exint3", "block0", "block3", "kcal0", "kcal3")
df_clean = na.omit(df[,c(baseline, exercise)])
```

```
corrplot.mixed(cor(df_clean),
               lower = "number",
               upper = "circle",
               tl.col = "black")
```

Question 2.3

```
library(jtools)
df_reg = df[,c("mortality", baseline)]
lm1 = lm(mortality ~ . , data = df_reg)
summary(lm1)
summ(lm1)
```

Question 2.4

```
df_reg2 = df[,c("mortality", baseline, exercise)]
lm2 = lm(mortality ~ . , data = df_reg2)
summary(lm2)
summ(lm2)
```