Dongyang Wang
Professor Taeb
Stat 528
2/6/2023

HW3

Q1

Since there are no extra feedback from Zhen, I will simply attach my answers in HW2 without modification in the following pages.

Dongyang Wang
Professor Taeb
Stat 528
1/26/2023

An Exploratory Data Analysis: Associations of the CHS Dataset

## Part I: Introduction

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 Siscovick's 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.

## Part II: EDA and Trend Plotting

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 number of calories expended does decreases. Furthermore, for everyone, 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.

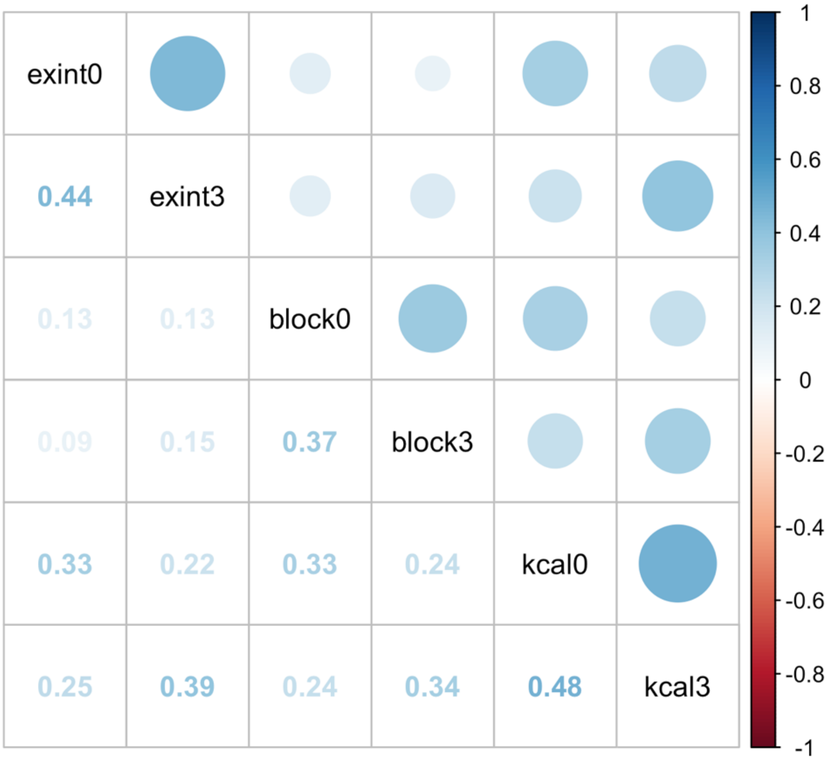


Figure 1: Correlation Plot between Exercise Variables

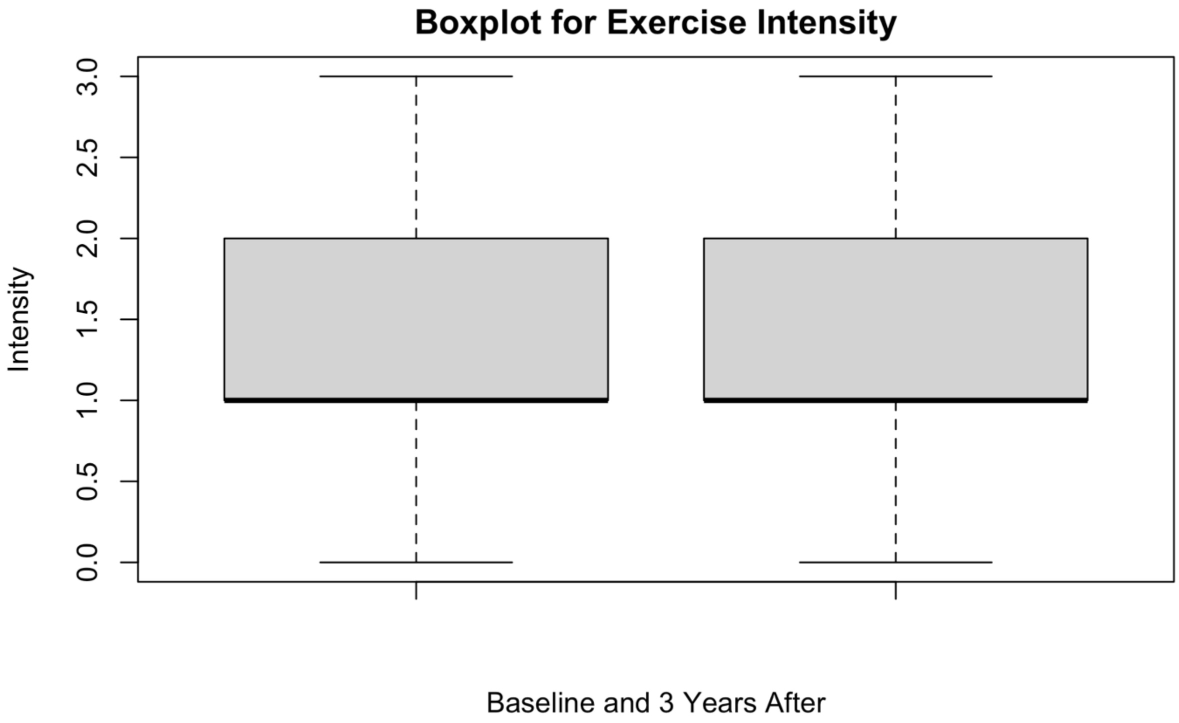


Figure 2: Boxplot for Exercise, baseline and 3 years after

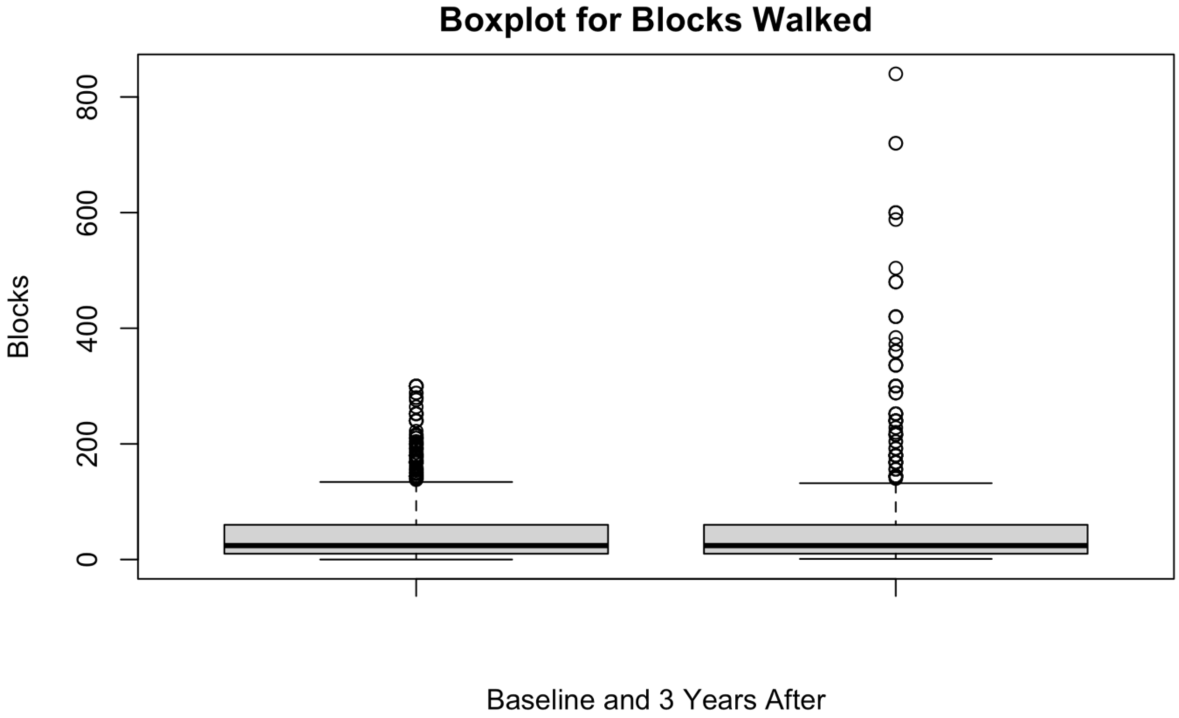


Figure 3: Boxplot for Blocks Walked, baseline and 3 years after

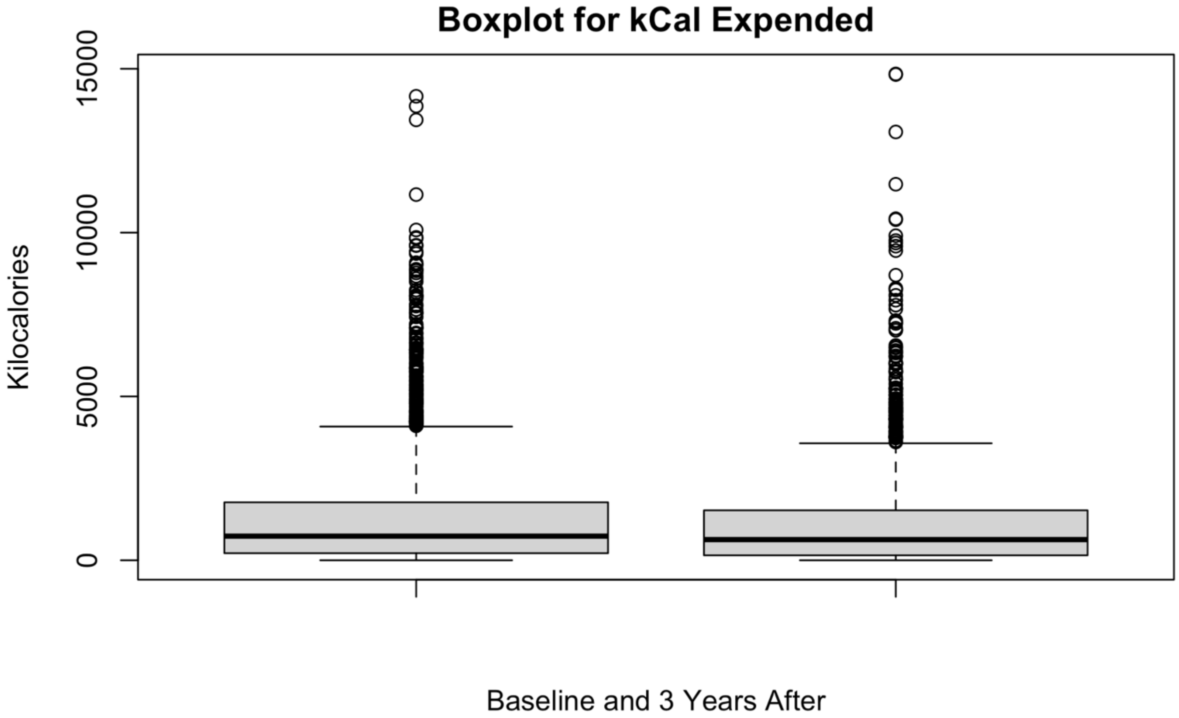


Figure 4: Boxplot for kCal expended, baseline and 3 years after

| | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | NA's |
|---|---|---|---|---|---|---|---|
| Exercise Intensity | -3.00 | 0.00 | 0.00 | -0.06 | 0.00 | 3.00 | 283.00 |
| Blocks Walked | -288.00 | -24.00 | 0.00 | 5.22 | 24.00 | 696.00 | 502.00 |
| kCal Expended | -12262.50 | -675.00 | -67.50 | -211.59 | 314.88 | 14531.00 | 289.00 |

Table 1 Table for Summary Statistics of the Exercise Variables, baseline and 3 years after

Additionally, a mosaic plot helps showing the trend for the changes in exercise intensity. It appears that most people remain unchanged in terms of their exercise intensity despite aging.

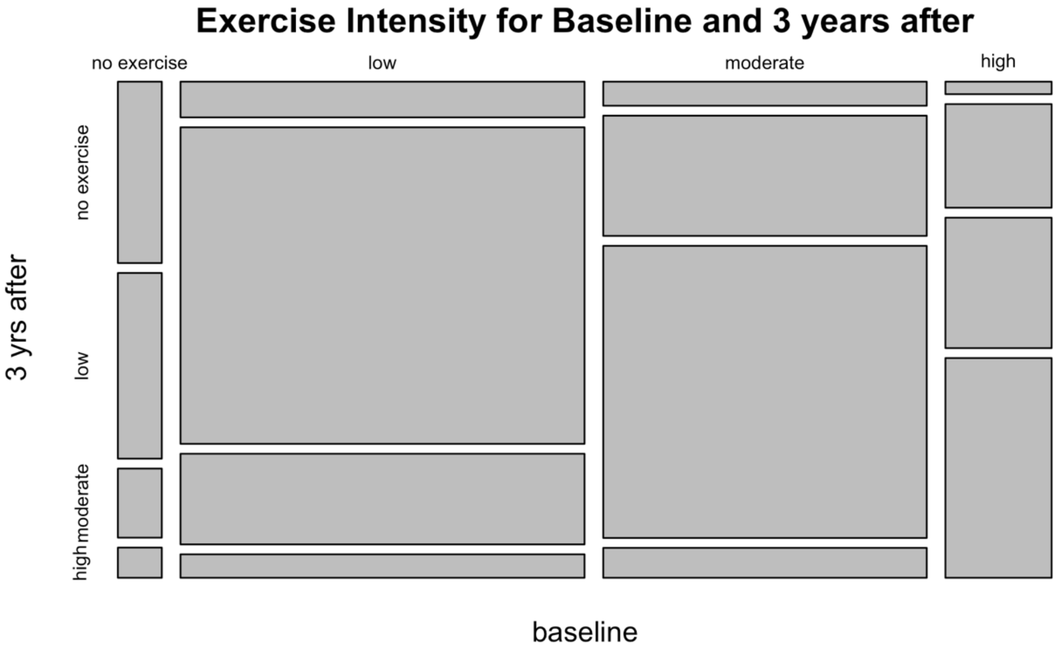


Figure 5: Mosaic Plot for Exercise Intensity, baseline and 3 years after

Before going into details about the baseline variables, 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.

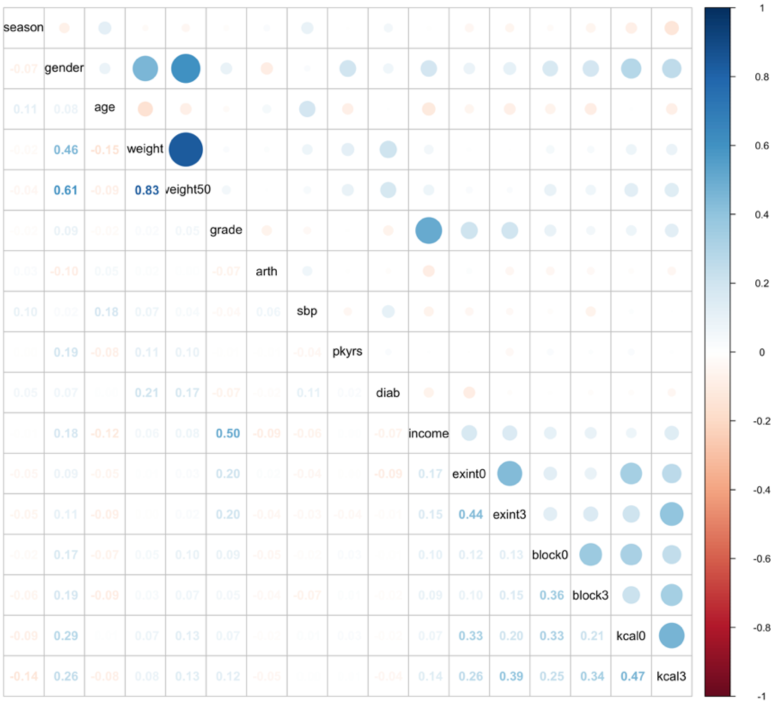


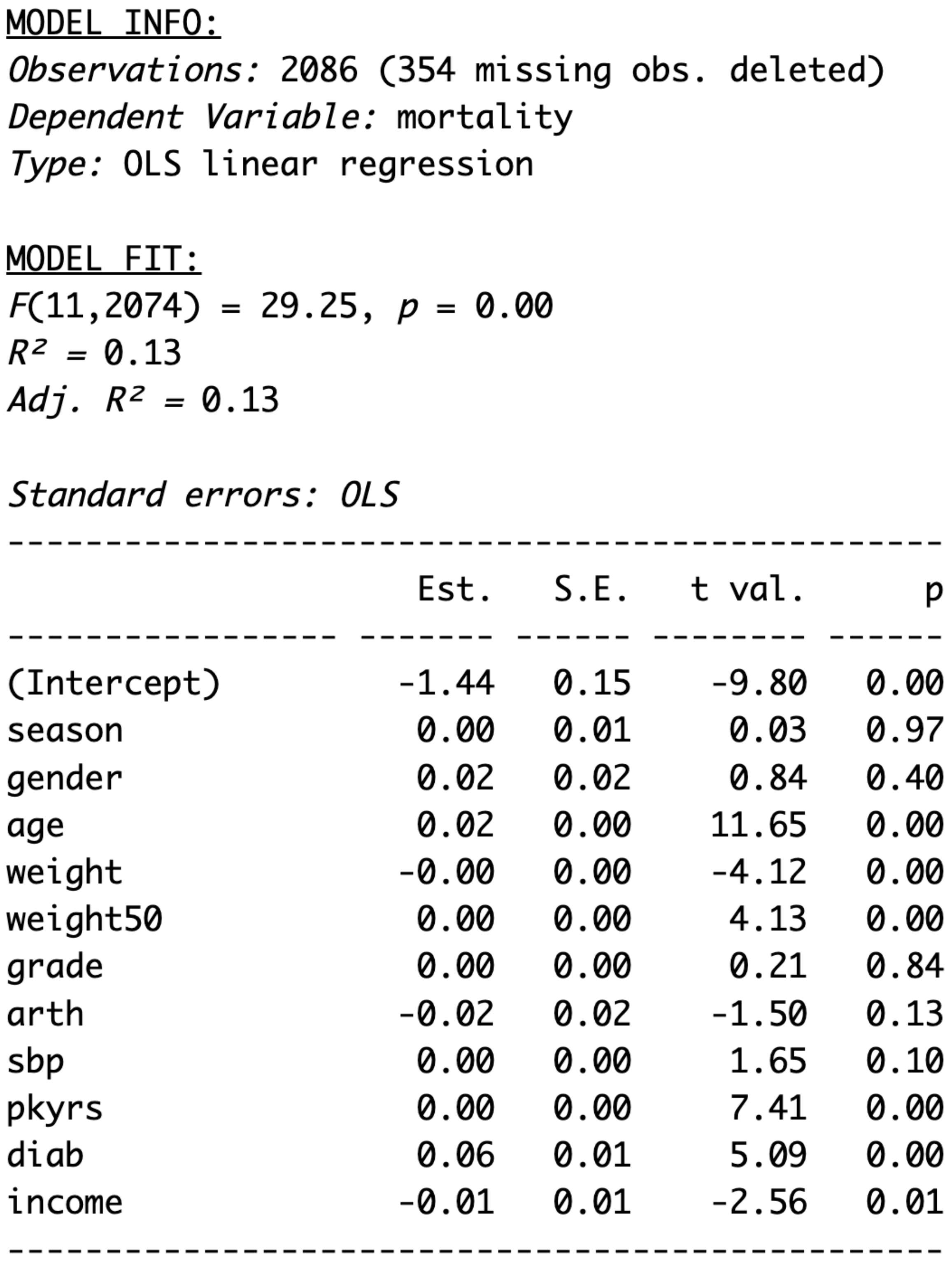
Figure 6: Correlation Plot between Exercise Variables and Baseline Variables

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.

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.

## Part III: Conclusion

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.

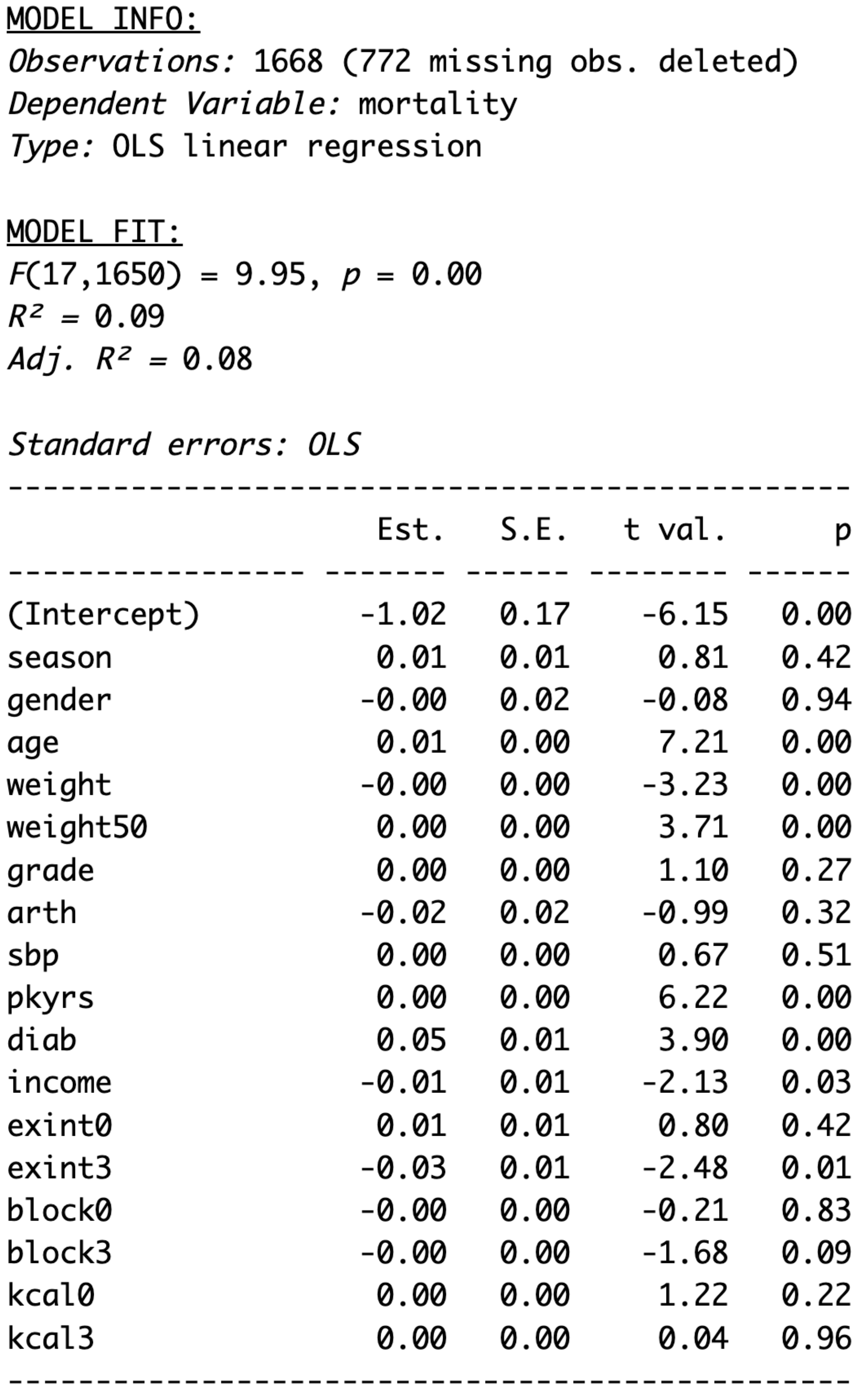


```
MODEL INFO:
Observations: 2086 (354 missing obs. deleted)
Dependent Variable: mortality
Type: OLS linear regression

MODEL FIT:
F(11,2074) = 29.25, p = 0.00
R² = 0.13
Adj. R² = 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
------------------------------------------------
```

Figure 7: Regression Output for Simple Linear Regression with Only Baseline Variables



```
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
R² = 0.09
Adj. R² = 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
------------------------------------------------
```

Figure 8: Regression Output for Simple Linear Regression with Both Baseline Variables and Exercise Variables

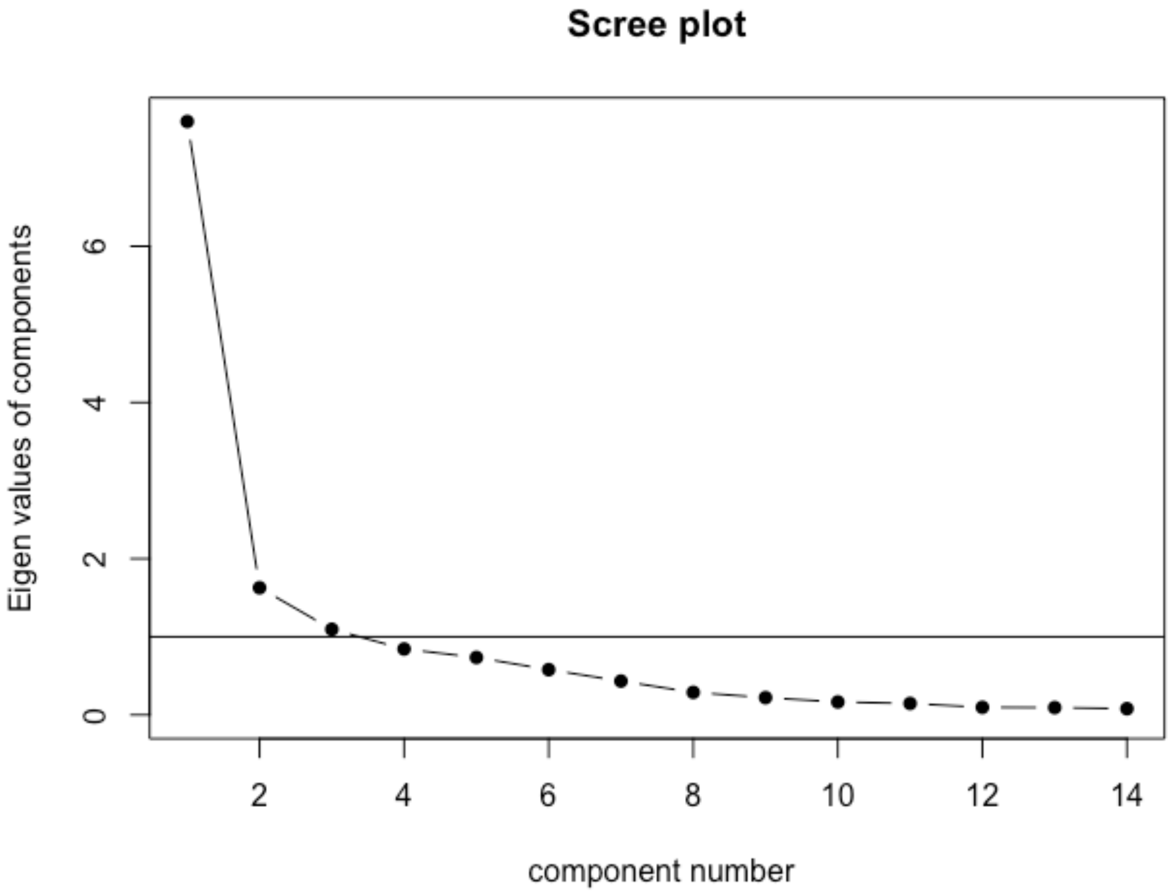
## References

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

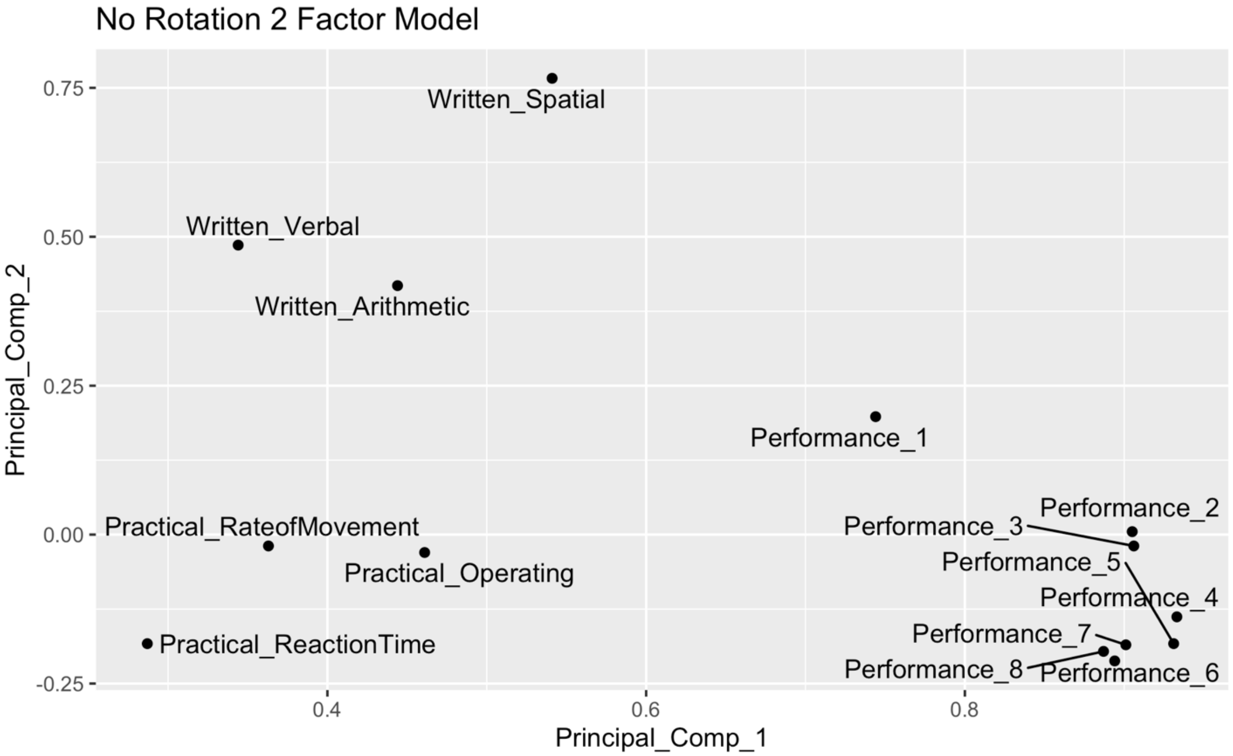
Siscovick, D.S. et al. (1997). Exercise intensity and subclinical cardiovascular disease in the elderly. *American Journal of Epidemiology*, **145**, 977–986.

Q2

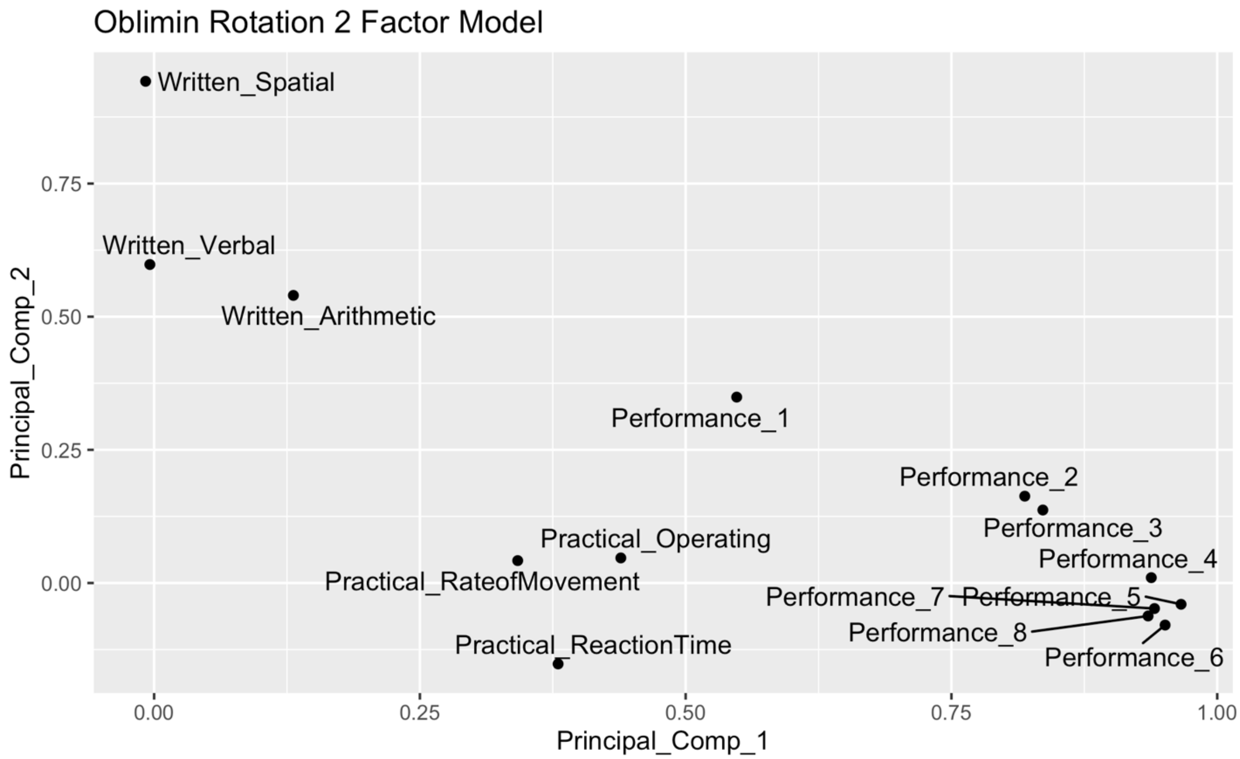
1. According to the web, “a scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always displays a downward curve.” ( https://www.theanalysisfactor.com/factor-analysis-how-many-factors/#:~:text=A%20scree%20plot%20shows%20the,be%20generated%20by%20the%20analysis. ) The elbow in this case shows that 2 factors should be used.



2. Based on the results, I have attached the plot as follows. The 14 variables have been decomposed to project on the 2 axes where there are 2 factors. It seems that the first component is capable of explaining the performance 1-8, where there is a small cluster on the right end of the x axis but their y value is negligible yet similar. Similarly, the written and practical skills are more or less clustered around the 0.4 of the x axis, but the y axis has more variation.



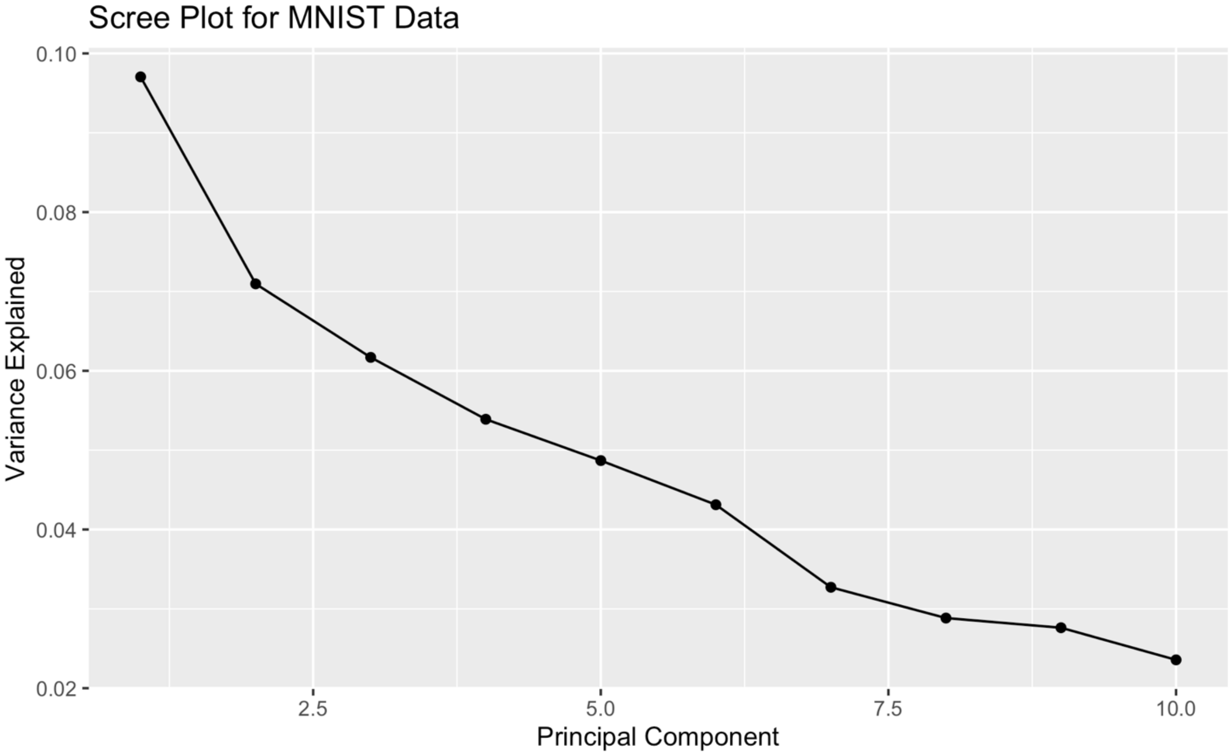
3. After the transformation, there appears approximately a downward linear relationship. The overall result is still similar to the previous question. But it seems that the clusters now are more evident, especially for the written abilities, practical abilities, and performance respectively. The relationship appears more organized and interpretable by the 2 factors.



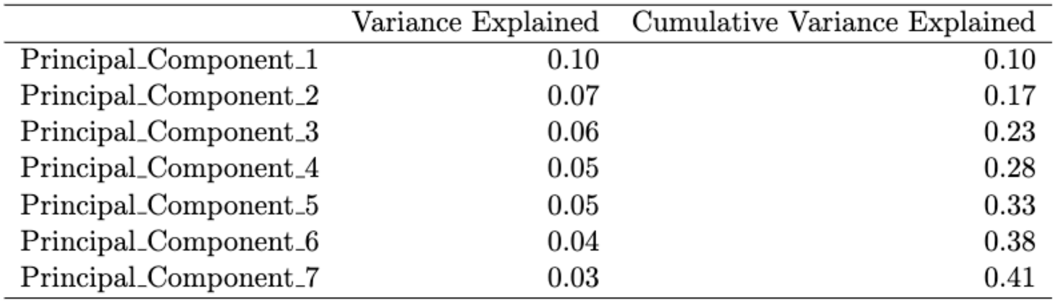
Q3.1

1. We should use the covariance matrix because in lecture we learned that If units of p variables are comparable, PCA based on covariance S may be more informative because units of measurement are retained. Since the handwritten digits are parsed out such each pixel is between 0 or 255, although some pixels in the beginning are mostly 0. But every single pixel has exactly the same standard for its value, so the use of covariance matrix is legit.

2. We generate a scree plot to show the principal components. The variance explained is generated by calculating the percentage of variance as a proportion of total variance. Plotting on variance explained, the following graph is obtained. In this case, we can say there are 7 principal components since there is approximately an elbow here. We can also possibly say there are 4 principal components, but since that the variance explained for each component is relatively small, 7 is probably better for prediction.

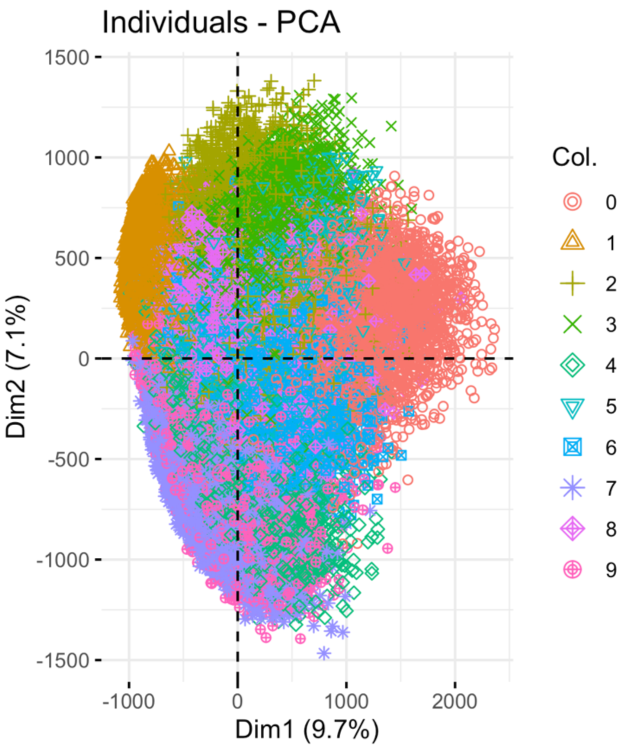


3. The following table displays the percentage of variance and cumulative percentage of variance explained by each principal component.



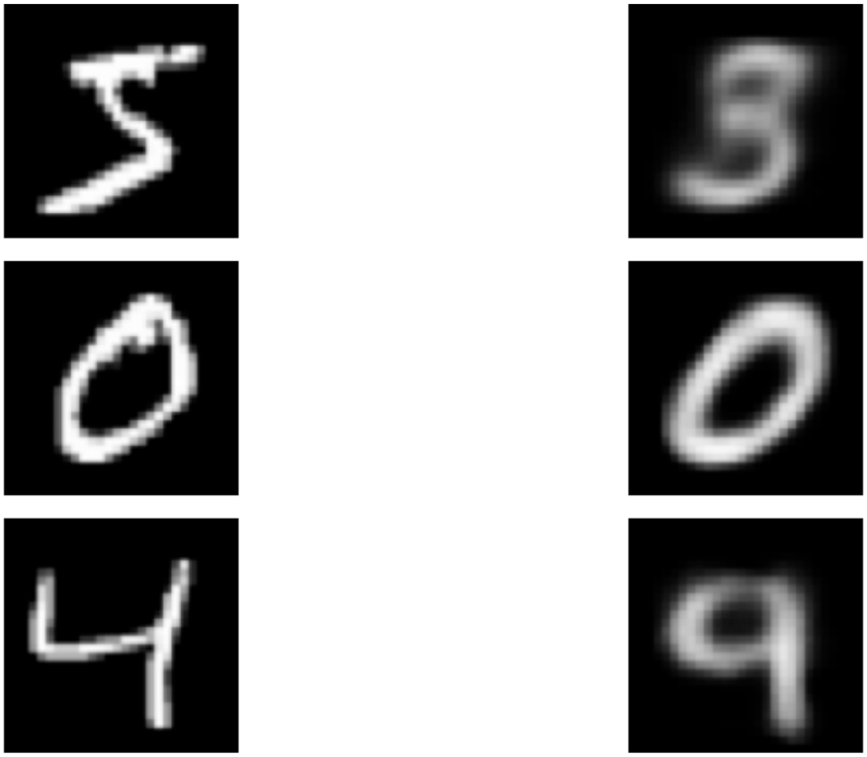
| | Variance Explained | Cumulative Variance Explained |
|---|---|---|
| Principal_Component_1 | 0.10 | 0.10 |
| Principal_Component_2 | 0.07 | 0.17 |
| Principal_Component_3 | 0.06 | 0.23 |
| Principal_Component_4 | 0.05 | 0.28 |
| Principal_Component_5 | 0.05 | 0.33 |
| Principal_Component_6 | 0.04 | 0.38 |
| Principal_Component_7 | 0.03 | 0.41 |

4. The projection on two axes appears a bit messy. The pattern is that the data points are pretty much around the centers based on the two principal components and different labels seem to revolve around different centers.

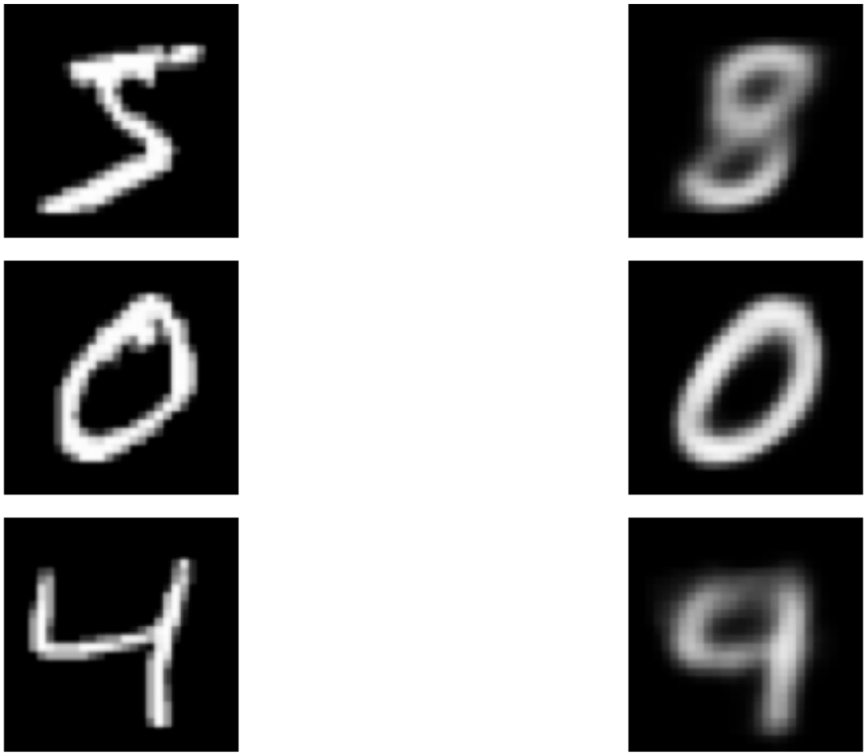


Q3.2

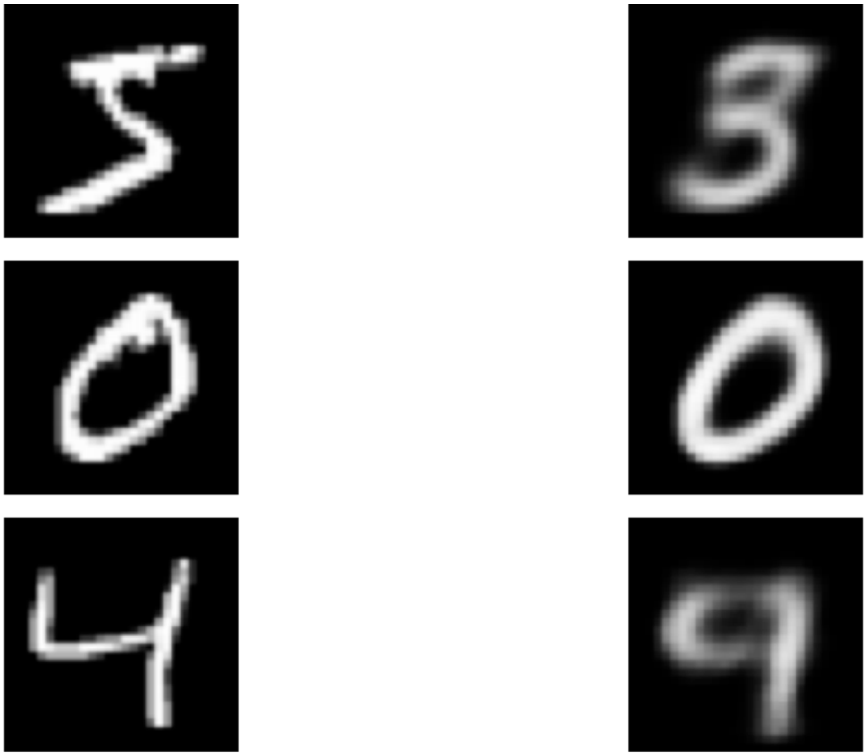
1. After training different VAEs with number of latent variables h = 2,3,4, I have obtained the following images. These comparisons were made on the first 3 images of the training set.



The above is the 2 latent factor model result.

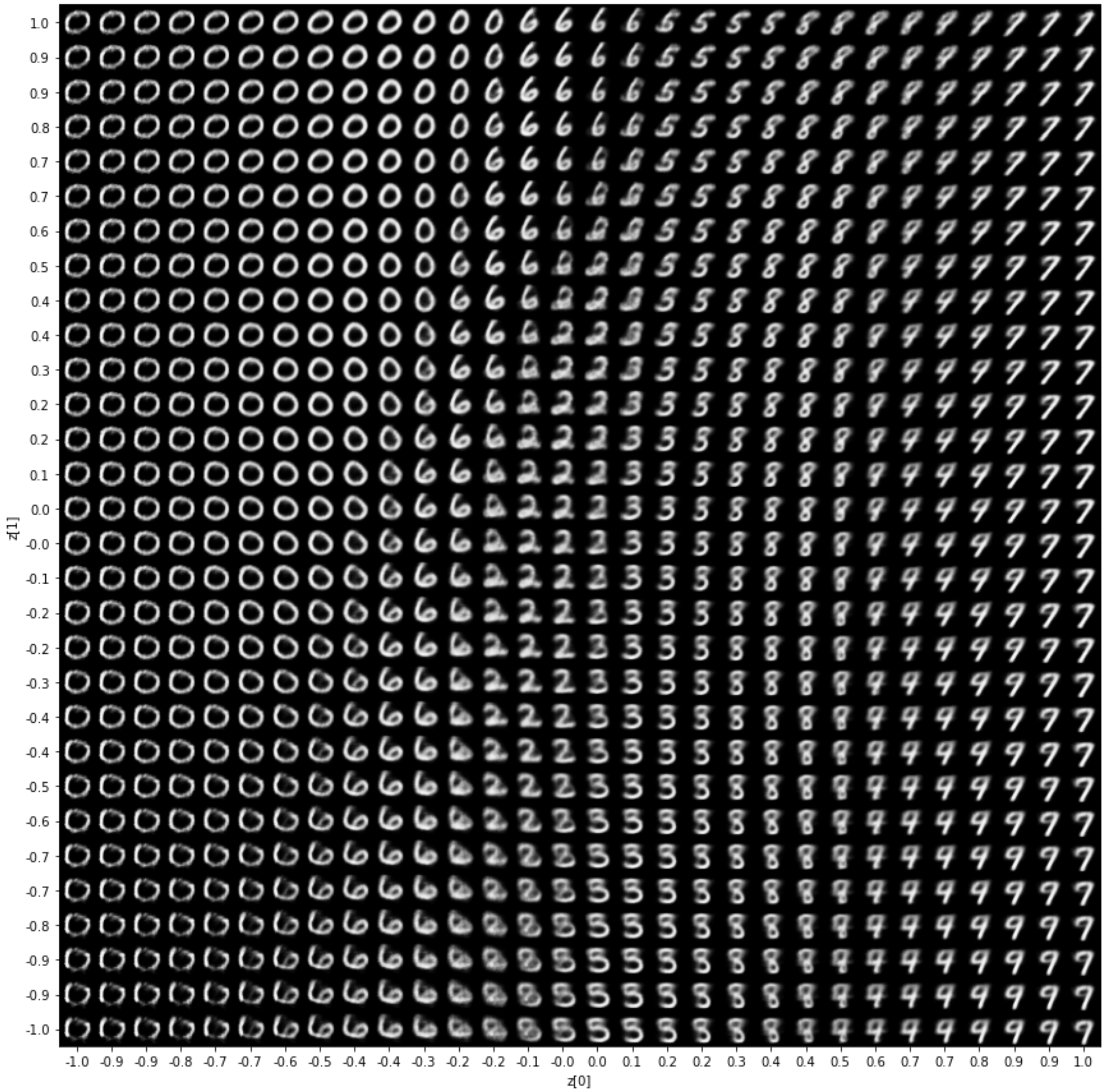


The above is the 3 latent factor model result.



The above is the 4 latent factor model result.

2. Based on latent traversals. The first latent variable seems the O shape of the digit. The second latent variable is more like the tail of 4 and 9. The third latent variable is more or less like the narrow or wide neck of the digit 8. The fourth latent variable is the straight or wiggly lines on the right side for 9, 3, 1. Pic source(https://keras.io/examples/generative/vae/)



3. I could not figure out this problem, but I think I understood the concept. Instead of putting gibberish here I am confessing my ingnorance in exchange for some mercy for partial credits.