CATEGORICAL DATA ANALYSIS OF CARDIOVASCULAR DISEASE

STAT-6395 Final Project

PROFESSOR: XINLEI (SHERRY) WANG

STUDENT: YICHIEN CHOU, ZHERUI LIN

SOUTHERN METHODIST UNIVERSITY

1. Introduction

Under the special background of the global epidemic, heath and diseases have become the most popular topic in the world. According to the CDC[1] (Centers for Disease and Control and Prevention): “People of any age who have serious underlying medical conditions might be at higher risk for severe illness from COVID-19.” Those underlying medical conditions include people who have serious heart conditions. Cardiovascular disease (CVD) is a general term that refers to the diseases that involve the heart or blood vessels. Some typical types of cardiovascular diseases known as a stroke and heart failure. According to the Harvard Health Blog[2]: “About 10% of patients with pre-existing cardiovascular disease (CVD) who contract COVID-19 will die, compared with only 1% of patients who are otherwise healthy.” So, it is a disease that we should pay special attention to.

The goal of this paper is to predict the presence or absence of cardiovascular disease with factors like age, weight, blood pressure, etc. By examined the relationship between CVD and the determine factors, we hope it will raise people’s health awareness.

2. Data Preprocessing

Data Description

The data source of this paper can be found on Kaggle: (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset?select=cardio_train.csv). It consists of 70000 rows and 13 columns. The ratio for the presence and absence of CVD is almost 1:1, so it’s perfect for modeling. But before inputting everything to the models, we will conduct some level of preprocessing to make sure the data is ready to use.

Table 1, variable description

Name	*Description*
Age	Days since born
Height	Height in centimeters
Weight	Weight in kilogram
Gender	1:women 2 : man
Ap_hi	Systolic blood pressure
Ap_lo	Diastolic blood pressure
Cholesterol	1: normal, 2: above normal, 3: well above normal
Gluc	1: normal, 2: above normal, 3: well above normal
Smoke	0: Not smoke 1: Smoke

[1] https://www.cdc.gov/coronavirus/2019-ncov/hcp/underlying-conditions.html

[2] https://www.health.harvard.edu/blog/how-does-cardiovascular-disease-increase-the-risk-of-severe-illness-and-death-from-covid-19-2020040219401

Alco	0:Not drinking Alcohol 1: Drinking Alcohol
Active	0: No Phyical Activity 1:Phyical Activity
Cadio	0: Absence of cardiovascular Disease 1: Presence of cardiovascular Disease

Feature Engineering

The first thing we did was looking for missing values. Fortunately, there was no sign of missing values. Then, we move to the data cleaning process. When we were looking into every feature, we find that the first column: ID is a list of numbers. It did not provide any related information with the CVD. So, we removed this column from the data.
The next thing needed to change is the age column. Oddly, the value of is five-digit numbers. It was recording the days since born, so we divide the values with 365 and keep all the decimals.

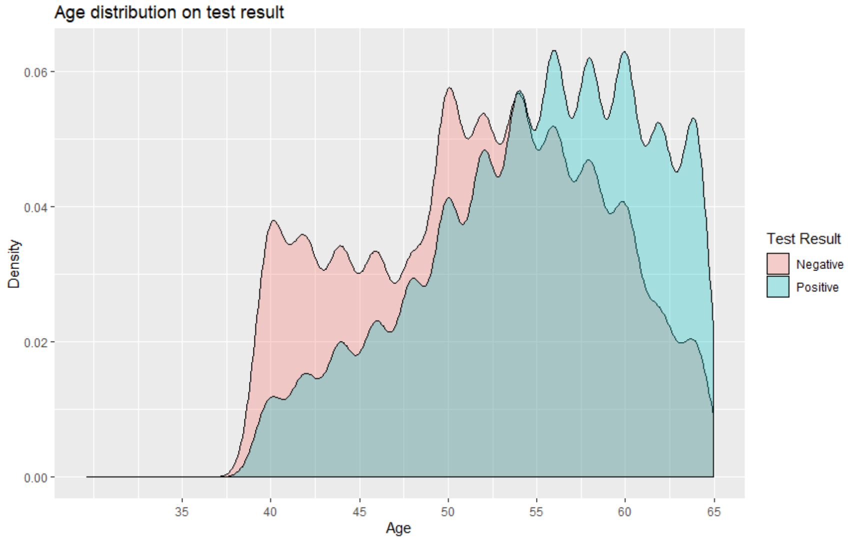
Then, when we researched the height and weight, we realized there's a potential relationship that exists between them. So, we used the Body Mass Index (BMI) to recalculate and transform them. The BMI calculates as $\frac{Mass\ (KG)}{Height^{2}(M)}$, by plugin the formula, we created a new variable named bmi.

Last but least, when we check the class of the variables, we found that all the variables are marked as an integer or numeric. We will not able to conduct a logistic regression if we keep it that way. Then, we transform some variables like gender into factors variables and correctly labeled them.

3. Exploratory Data Analysis

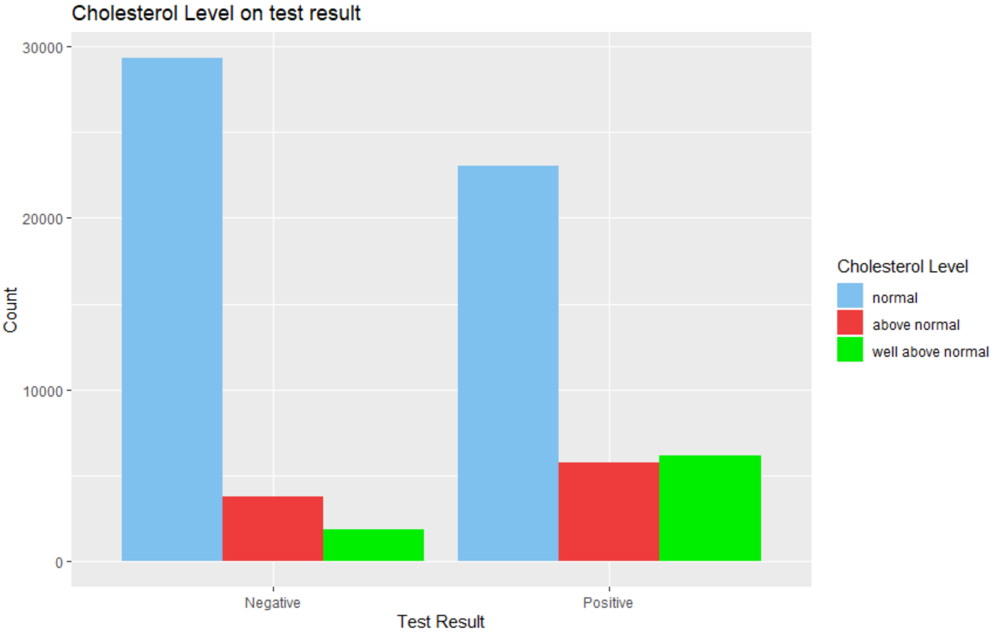
In this section, we will examine the relationship between CVD and 3 variables (age, cholesterol, mi). We mark the presence of CVD as "positive" and the absence as "negative". Hopefully, the result of finding may bring some insights to the people.

Figure 1, distribution plot of age on test result



The first variable we exam was Age. By build a density plot of age over the CVD test result, we find that the positive cases and negative cases overlap through age 35 to 65. However, when the age is below 55, the density of the negative cases is significantly higher than the positive cases. When the age is great 55, the density of positive is significantly higher than the negative cases. Age plays a big role in resulting CVD; the younger person is less likely to have CVD. Also, Age 55 is a cutoff point for the presence of CVD, that's where most positive and negative cases separate.

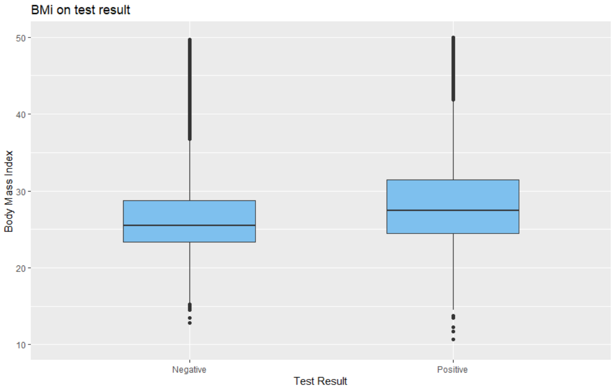
Figure 2, bar plot of cholesterol level on test result



The second variable we researched was the cholesterol level. Since the count of positive and negative is almost the same, building a bar plot will show us the difference in negative cases and positive cases on each cholesterol level. As we can see from the graph, the people who are test positive for CVD, their cholesterol level is much higher than the negative case. For a person whose cholesterol level is "well above normal", almost 80%

are having one or more CVD. That percent was only 60 % on the previous level. We can conclude that as the level of cholesterol increases, the presentence of CVD increases as well. A good practice to cut down the cholesterol level is to avoid too much sugar and carbohydrates on a daily diet. By consuming some oatmeal[3] is another way to lower the cholesterol level.

Figure 3, boxplot of BMI on test result



According to the CDC[4], the BMI from 18.5-24.9 is classified as normal or healthy wright and the BMI from 25.0-29.9 classify as overweight, 30.0 and above classify as obese. As we can see from the boxplot above, the person who does not have CVD, their BMI group around 25 which was the border on normal. For the person who have CVD, their BMI group around 27.5 which consider as overweight. So good control of the weight will help us to reduce the chance to have CVD. Exercise regularly is benefited to control the weight. Also, it's not only good for our body, mentally, it can also help us to reduce stress.

Table 2, frequency table

		Cardiovascular disease		
Features		**Negative**	**Positive**	**Total**
Gender	Women	22914	22616	45530
	Man	12107	12363	24470
Cholesterol	Normal	29330	23055	52385
	Above Normal	3799	5750	9549
	Well Above Normal	1892	6174	80666
Gluc	Normal	30894	28585	59479
	Above Normal	2112	3078	5190
	Well Above Normal	2015	3316	5331
Smoke	No	31781	32050	63831
	Yes	3240	2929	6169
Alco	No	33080	33156	66266
	Yes	1941	1823	3764
Active	No	6378	7361	13739

[3] https://www.health.harvard.edu/heart-health/11-foods-that-lower-cholesterol

[4] https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

	Yes	28643	27618	56261

When it comes to healthy, smoking always has a bad influence on our impression. So we build a 2 by 2 table to check the relationship between smoke and CDV. From the result of the table, the sample odds for smoking is 0.475, and nonsmoking is 0.502. The odd ratio is 0.946 which means the estimated odds of CDV in the smoking group to be 0.946 times odds to the nonsmoking group. The result suggests smoking negatively associated with CDV. This is a shocking result, but it does not mean smoking will not affect you in other ways.

4. Logistic Regression

In this section, we applied a logistic regression model to find the relationships between the presence of cardiovascular disease and other predictors.

Model Selection

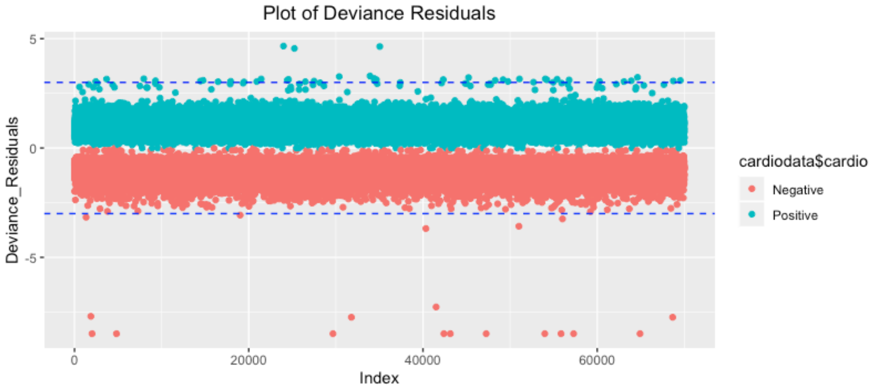
Our primary goal is to find the interpretation of the relationships between the absence of cardiovascular disease and other predictors from the logistic regression model, we aim to build a simpler model which is easier to interpret. We select the model by using three automated selection methods: forward, backward, and stepwise selection. The results were the same. The selection methods all kept all of the predictors.

Model Checking

In this section, we would check if the model fits well.

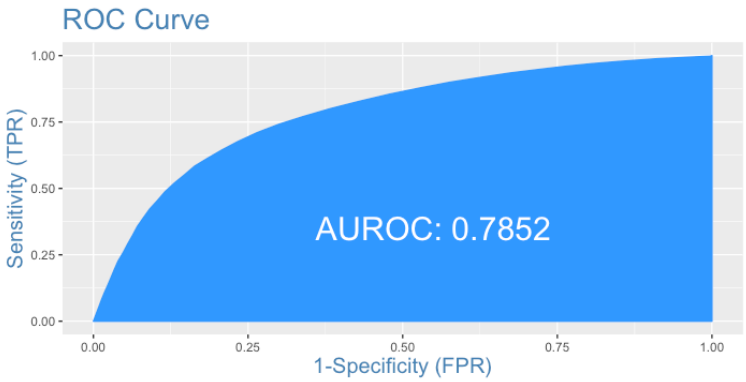
Since it is a balanced data and at least 10 outcomes of each type for every predictor, we did not need to remove the predictors due to unbalanced data. Therefore, we built a model using all the predictors (except BMI due to multicollinearity). There were no predictors that had the variance inflation factor (VIF) larger than 4, so we assumed the absence of multicollinearity. Furthermore, we used residual plot to see if the model fits well.

Figure 4, residual plot



From the Figure 4, we can see most of the observations fall within 3 and -3. Thus, there is no evidence of lack of fit.

Figure 5, ROC curve and AUC



The ROC is 0.7852, which means the model has good predictive power. The sensitivity in this model is 0.6762. The fraction of those with the cardiovascular disease correctly identified as positive by the test is 67.62%. The specificity in the model is 0 .7665. The fraction of those without the cardiovascular disease correctly identified as negative by the test is 76.65%.

Odds Ratio

In this part, we only mentioned the statistically significant (p<0.05) predictors:

(1) **Age**: The odds of the presence of cardiovascular disease increases by 5.5% for every 1 year older in age.
(2) **Height**: The odds of the presence of cardiovascular disease decreases by 0.6% for every 1 cm increase in height.
(3) **Weight**: The odds of the presence of cardiovascular disease increases by 1.5% for every 1 kg increase in weight.
(4) **Systolic blood pressure**: The odds of the presence of cardiovascular disease increases by 4% for every 1-unit increase in weight systolic blood pressure.
(5) **Diastolic blood pressure**: The odds of the presence of cardiovascular disease increases by 0.3% for every 1-unit increase in weight diastolic blood pressure.
(6) **Cholesterol**: Compared to the patients with a normal volume of cholesterol, the odds of the presence of cardiovascular disease increases by 52.5% if the patients had a higher volume of cholesterol in the body; The odds of presence of cardiovascular disease increases by 218.8% if the patients had a volume of cholesterol well above normal in the body.
(7) **Glucose**: Compared to the patients the with level of glucose above normal in blood, the odds of the presence of cardiovascular disease decreases by 28.8% if the patients with the level of glucose well above normal in the blood.
(8) **Smoke**: Compared to the non-smokers, the odds of the presence of cardiovascular disease for smokers decreases by 12.3%.
(9) **Alcohol intake**: Compared to the patients without alcohol intake, the odds of the presence of cardiovascular disease for patients with alcohol intake decreases by 15.6%.
(10) **Physical activity**: Compared to the patients having physical activity, the odds of the presence of cardiovascular disease for the patients did not have physical activity

decreases by 19%.

5. Conclusion

From the EDA section, the first thing we find out is that smoking is not positively related to the CDV. The second thing we find is that age is a strong indicator of the presence of CDV. Below the age of 55, the person who does not have CDV is a lot more the ones who do. However, as soon as the age past 55, We see huge growth in the number of positive cases. The Third thing we researched was the cholesterol level. A person whose cholesterol level is above or well above normal seems more like to have CDV. The last thing we looked at was the BMI. By the BMI standard, when a person classified as normal. There are less likely to have CDV.

From the odds ratio section, we can find that as the age or weight increases, the probability of the presence of cardiovascular disease increases. Furthermore, the increasing of systolic or diastoc blood pressure would also increases the probability of the presence of cardiovascular disease. Surprisingly, we found that as height increases, the probability of the presence of cardiovascular disease decreases.

For categorical predictors, the results were totally out of the blue. The group of glucose above normal in blood seems has the higher probability of the presence of cardiovascular disease in comparison with the group of glucose well above normal in the blood. What's more, smokers seems has a lower probability of cardiovascular disease compared to non-smokers. The surprising results also happened to the patients with alcohol intake, the patients with low physical activity.

In general, people thought that smoking, drinking, and lack of physical activity might lead to a high risk of cardiovascular disease. However, the outcome of odds ratios was just the opposite. The reason why it happened might because we did not take the interaction term into account.

Reference

[National Center for Immunization and Respiratory Diseases, 2019]: https://www.cdc.gov/coronavirus/2019-ncov/hcp/underlying-conditions.html

[Dara K. Lee Lewis, 2020]: https://www.health.harvard.edu/blog/how-does-cardiovascular-disease-increase-the-risk-of-severe-illness-and-death-from-covid-19-2020040219401

[Harvard Medical School, 2019]: https://www.health.harvard.edu/heart-health/11-foods-that-lower-cholesterol

[Center for disease control and prevention, 2020]: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

[GraphPad]: https://www.graphpad.com/guides/prism/7/statistics/stat_interpreting_results_contingen_2.htm

Appendix1: Individual members' time and effort allocation

Zherui Lin: I was responsible for the introduction, data preprocessing, and the EDA Part. Also, we wrote the first part of the conclusion. The way we allocate this project was I would be focusing on everything before building the model and Yi-Chien will take care of the rest. By doing that, we should get about the same amount of work. I ended up using around 11 hours to finish my parts. The most difficult part was looking relationship between variables. And I had to switch back and forth to decide what kind of plots will best represent the relationship. But, I think Yi-Chien spent about the same amount of time with me, o it's a great team effort overall.

Yichien Chou: I was responsible for the logistic regression section and part of the conclusion. First, I built a logistic regression model, and I did the process of model selecting, model checking, and model comparsion (I used the Hosmer-Lemeshow goodness of fit test, check if the two-way interaction term exists, and whether can we drop the main effect, but I did not mentioned the above processes in paper due to page limit). The time I spent on it was about 10 hours or more. Zherui did a great job in data pre-processing, so that I could analysize the data directly. He also gave a detailed introduction of cardiovascular disease.

Appendix2: R code

```
#read data
cardiodata<-read.csv("/Users/jason13nn/Desktop/SMU/Spring 2020/STAT 6395 (CDA)/Final project/cardio_train.csv",sep = ";")
head(cardiodata)

# remove ID
cardiodata$id<-NULL

#transform days into years for Age
cardiodata$age<-cardiodata$age/365

str(cardiodata)

# factorize variables
cardiodata$gender<-factor(cardiodata$gender,labels=c('women','men'))
cardiodata$cholesterol<-factor(cardiodata$cholesterol,labels=c("normal","above normal","well above normal"))
cardiodata$gluc<-factor(cardiodata$gluc,labels=c("normal","above normal","well above normal"))
cardiodata$smoke<-factor(cardiodata$smoke,labels = c("No","Yes"))
cardiodata$alco<-factor(cardiodata$alco,labels = c("No","Yes"))
cardiodata$active<-factor(cardiodata$active,labels = c("No","Yes"))

# Target Variable
cardiodata$cardio<-factor(cardiodata$cardio,labels = c("Negative","Positive"))
str(cardiodata)

# Feature enginering
library(ggcorrplot)

#add new varable BMI
cardiodata$bmi<-(cardiodata$weight)/((cardiodata$height/100)^2)

#check missing value
sapply(cardiodata, function(x) sum(is.na(x)))
# No missing values

# Looking into factor variables first
ggplot(cardiodata, aes(x=cardiodata$age, fill=cardiodata$cardio)) + geom_density(alpha=.3)+
  labs(title="Age distribution on test result",
       x="Age",y = "Density") +
  scale_fill_discrete(name = "Test Result",labels = c("Negative", "Positive"))+
  scale_x_continuous(breaks=seq(35, 70, 5))
# older people easier to have cardio

#cholesterol
ggplot(data = cardiodata, mapping = aes(x = cardio, fill = cardiodata$cholesterol)) +
  geom_bar(stat = 'count', position = 'dodge')+
  labs(title="Cholesterol Level on test result",
       x="Test Result",y = "Count",fill="Cholesterol Level")+
  scale_fill_manual(values=c("skyblue2","brown2","green2"))

# BMI
ggplot(data=cardiodata, aes(x=cardio,y=bmi)) +
  geom_boxplot(width=0.5,fill = "skyblue2")   +ylim(10,50)+
  labs(title="BMi on test result",
       x="Test Result",y = "Body Mass Index")
  theme(plot.background = element_blank(),
          panel.background = element_blank(), axis.line = element_line(size=3, colour = "black"),
          axis.text.x=element_text(colour="black", size = , angle = 40, hjust=1),
          axis.text.y=element_text(colour="black", size = 6))

#logistic regression
cardio.lg <- glm(cardio ~., cardiodata[,-13], family = "binomial")
summary(cardio.lg)

#check multicollinearity
library(car)
vif(cardio.lg)
#no VIF larger than 4.

#Automated backward selection using AIC
#specify the full model crabs.fit1 to start
step(cardio.lg, direction="backward")

#Automated forward selection using AIC
#specify the null model to start
cardio.null <- glm(cardio ~1, family=binomial, data=cardiodata[,-13])
#specify the full model crabs.fit1 to stop
step(cardio.null, scope=list(lower=cardio.null, upper=cardio.lg), direction="forward")

#Automated stepwise selection using AIC
step(cardiodata[,-13], scope=list(upper=cardio.lg), direction="both")

#ROC
library(InformationValue)
```

```
cardio <- ifelse(cardiodata$cardio=="Negative",0,1)
cardio <- as.factor(cardio)
lg.predict <- predict(cardio.lg, cardiodata, type="response")
ROC <- plotROC(actuals = cardio, predictedScores = lg.predict)
ROC

#sensitivity
sensitivity(cardio, lg.predict)

#specificity
specificity(cardio, lg.predict)

#Residuals plot
Index <- 1:dim(cardiodata)[1]
# deviance residuals
Deviance_Residuals <- residuals(cardio.lg)
dff <- data.frame(Index,Deviance_Residuals,cardiodata$cardio)

ggplot(data = dff, mapping = aes(x = Index,y = Deviance_Residuals,color = cardiodata$cardio)) +
  geom_point() +
  geom_hline(yintercept = 3,linetype = "dashed", color = "blue") +
  geom_hline(yintercept = -3,linetype = "dashed", color = "blue") +
  labs(title = "Plot of Deviance Residuals") +
  theme(plot.title = element_text(hjust = 0.5))
#Most of the residuals falls with (-3, 3).

#Hosmer-Lemeshow test
#library(ResourceSelection)
#hoslem.test(cardiodata$cardio, fitted(cardio.lg), g = 10)

#interaction term
#cardio.lg.2 <- glm(cardio ~ (.)^2, cardiodata[,-13], family = "binomial")
#summary(cardio.lg.2)
#anova(cardio.lg.2, cardio.lg, test="Chisq")

#See whether we can drop any main effect
#drop1(cardio.lg, test="Chisq")

summary(cardio.lg)
```