ABSTRACT:
In this study, a survey data collected by WHO with 69269 records was used to analyze the relationships between life habits and cardiovascular disease. The logistic regression model can be applied to fit the data with binary outcome and 6 covariates including age, BMI, gender, smoking status, alcoholism status, regular activities indicator. Based on the BIC criteria and backward model selection process, a final logistic regression model containing all covariates of interest was fitted. The model diagnostics and comparison methods showed that the logistic regression model fits data well with less high influential points and outliers. The conclusions are robust and consistent in different logistic regression models. The results suggested that non-smoker ($OR = 0.93, P = 0.0388$), regular physical activities ($OR = 0.84, P < 0.001$) are significantly associated with the decreased odds of presence of cardiovascular diseases. However, the alcoholic intake ($OR = 0.94, p = 0.0875 > 0.05$) shows no significant associations for the cardiovascular diseases.

1 Introduction

Cardiovascular disease (CVD) is the most common disease of heart and blood vessels. Researchers have shown that CVD can be diagnosed with clinical testing and history records. Some observational studies can be used to support the role of changing life-style related risk factors such as keeping diet, physical exercises and alcohol consumption in CVD prevention. [1] A health survey conducted by WHO has collected the relative data of more than 60,000 people to analyze the association of CVD and life habits including smoking, physical activities and alcohol usage. In this project, the survey data can be acquired in Kaggle. The primary purpose of this project is to test if there is any significant relationship between life habits and CVD by fitting logistic models. The covariates are: age (in years), gender, body mass index (BMI), smoking indicator, regular activities indicator and alcohol intake indicator. The response variable (binary outcome) is whether the person has CVD at present. The original data is complete but it contains some extreme values. By eliminating unreasonable sample with extreme BMI levels that larger than 105 or lower than 8 according to the world-wide records, the data with 69269 samples can be used to do modeling.

Table 1: Summary Statistic of Survey Data

Variables	Summary Statistics	Interpretation
age	[29,64], MEAN = 52.84	age (in years) when taking the survey
BMI	[8.01,86.78], MEAN = 27.49	body mass index calculated by height and weight
gender	FEMALE:MALE = 45029:24210	gender
smoke	SMOKE:NON-SMOKE = 6098:63171	who have smoking history (at least once before)
alco	YES: NO = 3731:65538	who is alcoholism (alcohol use disorder)
active	YES:NO = 55694:13575	who have regular physical activities
cardio	YES:NO = 34607:34662	presence of cardiovascular diseases

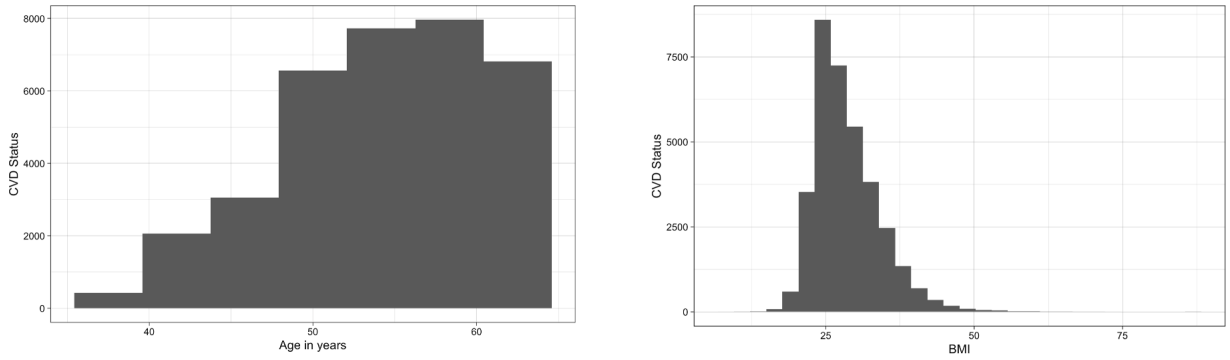

Figure 1: Histogram of the Number of CVD

The summary statistic of the collected data is shown in Table 1. In this survey data, the response variable cardio is balanced for two groups. For the continuous covariates age and BMI in Figure 1, it seems that the

number of CVD increases as age increases and it also shows a quadratic patterns between the number of CVD and BMI. In this case, the quadratic terms would be added into the logistic regression model to analyze the relationships between the covariates and presence of CVD.

2 Methods

2.1 Model Selection

From the exploratory data analysis, we will consider the full model with all quadratic terms for age and BMI, as well as all first-order term for each covariates. Then the model selection methods are applied to obtain the best model to analyze the main effect of life habits. The selection criteria would be based on the Bayesian information criterion (BIC): the smaller BIC, the less information lost by fitting the model. Compared with Akaike information criterion (AIC), BIC puts more penalty to the model since the sample size is large in this survey data. In order to fit a model which is suitable to explain the associations, BIC can be used as the model selection criteria. In this case, the model selection methods can be started with a full model that contains all first-order covariates and second-order terms for continuous covariates. Then by backward selection method, we will omit the covariate for the new selected model with least BIC. We will repeat this process until the BIC of new model is larger than the previous one. However, in order to check the relations between life habits and heart diseases, the categorical covariates related to smoking, alcoholic usage and physical activities are always kept in the model. Finally, the 'best' model with the smallest BIC can be used to analyze the association between cardiovascular disease and life habits.

2.2 Logistic Regression Model

In order to test if there is any significant relationship between CVD which is the binary outcome and life habits including smoking, physical activities and alcohol usage, the logistic regression models would be fitted with model selection methods to get the final model. The logistic regression model is:

$$\log(\frac{p}{1-p}) = \beta_0 + X\beta$$

where p is the probability of presence of cardiovascular diseases. X is the covariate matrix, the β_0 and β are corresponding intercept and regression coefficients. The logistic model will be fitted in every model selection steps. The BIC can be calculated by $BIC = \log(n)k - 2\log(L)$ where k is the number of covariates, n is the sample size and L is the likelihood. The likelihood ratio test was used for selecting significant risk factors associated with CVD. Given the significant level of 0.05, the covariate with p-value lower than 0.05 can be considered as statistically significant effect.

2.3 Model Diagnostics

In order to validate the statistical testing based on the previous logistic regression model, model diagnostics methods would be applied to check the model assumptions. Firstly, the Pearson's residuals and Deviance residuals can be calculated by the final model. A indication of lack of fit can be checked by box-plot of two types of residuals: if there is no lack-of-fit, the box-plots should be similar. Then the residuals plots will be also used to check the goodness of fit: if there is no lack-of-fit, the plots should show non-systematic pattern. Finally, the Runs test can be used to detect if there is any systematic pattern. The null hypothesis is H_0 : there is no lack-of-fit: if we reject H_0, then there is lack-of-fit in this model.
Finally, the leverage and Cook's distance will be calculated to check if there are some high influential points or outliers. A high leverage points with high Cook's distance might be a suspicious high influential points or outliers. If there is any influential point, the summary statistic of these points would be checked. If possible, they can be both removed and then refit the same final model with the rest of the data to check if there is any modification on the conclusion.

2.4 Model Comparison

Instead of adding quadratic term into the model selection methods, a potential model that discrete the BMI covariate might be available to fit the data. Based on the definition of healthy BMI for people, people with BMI lower than 18.5 might be considered as 'underweight', people with BMI higher than 25 can be seen as 'overweight'. People with BMI ranging between 18.5 to 25 can be denoted as 'normal'. These discrete BMI indicators can be added to the full model by replacing the original BMI covariate and finally acquired the final

model with the same model selection methods. In addition, from the results of model diagnostics, if there are many suspicious high influential points, the final model would be refitted without the high influential points to see if there is any change to the statistical inference. In this project, some potential models can be obtained to compare the statistical results.

3 Results

3.1 Model Fitted Results

Based on BIC selection criteria and backward selection methods, the best model with least BIC can be obtained:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{I[gender = MALE]} + \beta_3 \text{BMI} + \beta_4 \text{BMI}^2 + \beta_5 \text{I[smoke=NON-SMOKE]} + \beta_6 \text{I[alco=YES]} + \beta_7 \text{I[active=YES]}$$

where $BIC = 89548.4$. The estimated regression coefficients are shown in Table 2: it showed that non-smoker ($\hat{\beta}_5 = -0.0647, OR = 0.93, P = 0.0388$), regular physical activities ($\hat{\beta}_7 = -0.1718, OR = 0.84, P < 0.001$) are significantly associated with the decreased odds of presence of cardiovascular diseases. When fixing other covariates, male group tends to have higher odds compared with female group ($\hat{\beta}_2 = 0.163, OR = 1.18, p < 0.001$). However, since the estimated coefficient of alcoholic intake is not statistically significant ($\hat{\beta}_6 = -0.064, OR = 0.94, p = 0.0875 > 0.05$), we cannot conclude that the alcoholic usage would cause a significant effect on the presence of cardiovascular diseases. The regression coefficient of age is estimated as $\hat{\beta}_1 = 0.0697 > 0$ with p-value lower than given significant level 0.05. By fixing other covariates, it suggests that the advanced age tends to increases the log odds of the presence of cardiovascular diseases. The regression coefficients of BMI and its square term are both statistically significant, when fixing other covariates, the log odds is increases as the BMI increases from the minimum 8 to $BMI = 98.7$ and then the log odds will decreases as the BMI increases after $BMI = 98.7$. However, since the reasonable range of BMI is 8 to 87 approximately by world-wide records, the records with BMI higher than 98 are suspicious outliers, we can conclude that the higher BMI is statistically associated with the odds for the cardiovascular disease.

Table 2: Model Fitting Results of Logistic Regression Model

-	Estimate β	Standard Error	z-statistic	p-value
(Intercept)	-7.3451286	0.1559567	-47.097	< 0.001
age	0.0696584	0.0012213	57.034	< 0.001
gender(=MALE)	0.1629693	0.0178795	9.115	< 0.001
BMI	0.1918029	0.0092750	20.680	< 0.001
BMI^2	-0.0019430	0.0001477	-13.155	< 0.001
smoke(=NON-SMOKE)	-0.0647089	0.0313233	-2.066	0.0388
alco(=YES)	-0.0642710	0.0376110	-1.709	0.0875
active(=YES)	-0.1717960	0.0201185	-8.539	< 0.001

3.2 Diagnostics Results

The logistic regression model with the covariates discussed above can be validated by the following diagnostics results. For the goodness-of-fit of the logistics regression model, the box-plot for Pearson's residuals and Deviance residuals in Figure 2 [Right] shows that there is no indication of lack-of-fit for this selected model since two box plots are very similar even though the Pearson's residuals have larger range than the deviance residuals.

To further check the goodness-of-fit, the residuals plot in Figure 2 [Left]. The red smoothing line are fluctuated on 0 for both Pearson's residuals plot and Deviance residuals plot, which also indicates that there is no significant lack-of-fit. Compared with the model without quadratic term of BMI in Appendix-2, it is reasonable to add the quadratic term of BMI in the model. Finally results of the two-sided Runs test ($p = 0.152$) shows that there is no statistically significant lack-of-fit for the final logistic regression model. In this case, the selected model can fit the data well.

Finally, the influential points and outliers can be detected by the leverage and Cook's distance. For the 69269 survey records, there are 6757 high leverage points. The 6757 high leverage points (about 9% of the original data) contain all the records with alcoholism people, however the final selected model does not show significant effect of alcoholic usage to the cardiovascular disease. The high leverage points also contains all records with normal BMI (18.5-25). For these high leverage points, there are 3 leverage points with high

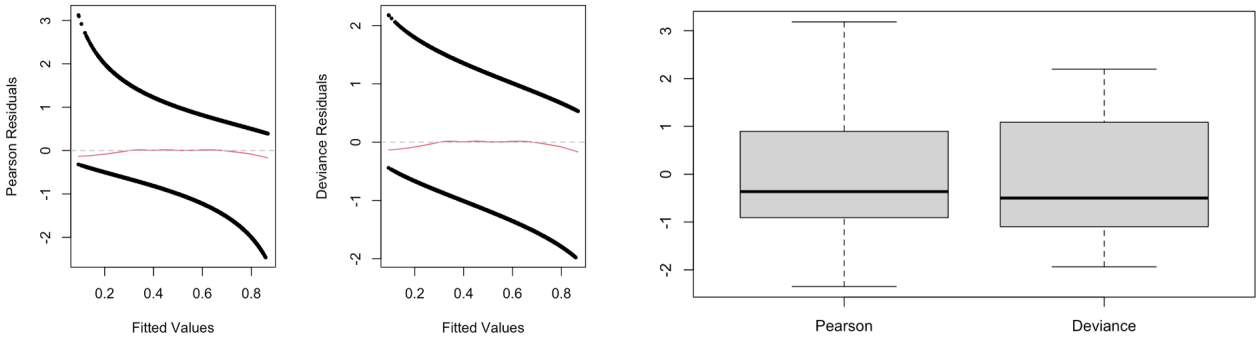

Figure 2: Residuals Plot [Left] and Box Plot for Residuals[Right]

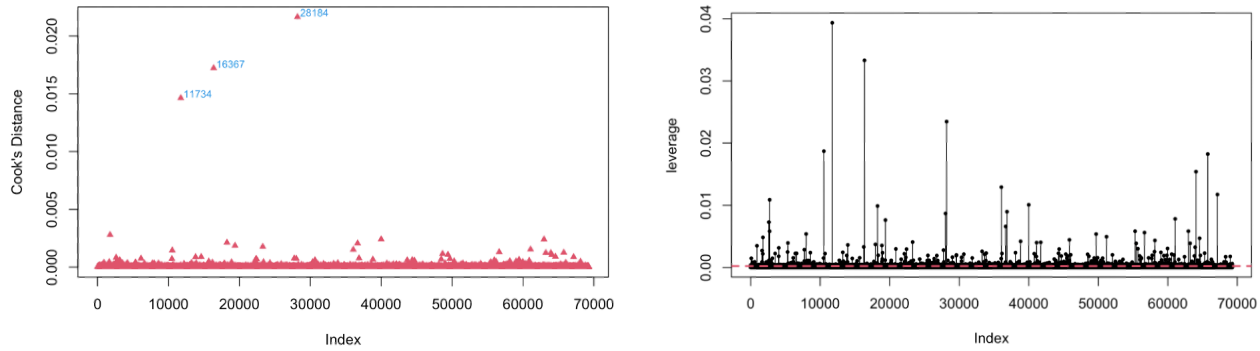

Figure 3: Cook's Distance Plots (Left) and Leverage Plots (Right)

Cook's distance which might be suspicious influential points and outliers. These points are with high value of BMI (all higher than 85, compared with the maximum $BMI = 86.78$ in the survey record).

3.3 Model Comparison Results

From the model diagnostics results, the new logistic regression model without the suspicious high influential points can be refitted to compare with the previous final model.

Table 3: Model Comparison: Estimate of β (p-value)

-	Final Selected Mode	Model Without Suspicious Points
(Intercept)	-7.3451286 ($p < 0.001$)	-7.4923655 ($p < 0.001$)
age	0.0696584 ($p < 0.001$)	0.0696169 ($p < 0.001$)
gender (=MALE)	0.1629693 ($p < 0.001$)	0.1623514 ($p < 0.001$)
BMI	0.1918029 ($p < 0.001$)	0.2021808 ($p < 0.001$)
BMI^2	-0.0019430 ($p < 0.001$)	-0.0021168 ($p < 0.001$)
smoke (=NON-SMOKER)	-0.0647089 ($p = 0.0388$)	-0.0644733 ($p = 0.0396$)
alco (=YES)	-0.0642710 ($p < 0.0875$)	-0.0643245 ($p = 0.0872$)
active (=YES)	-0.1717960 ($p < 0.001$)	-0.1717210 ($p < 0.001$)

The model comparison results can be shown in Table 3. The estimate of regression coefficients are very close for two models. For both two models, the increased age, smoker, male people without regular physical activities are significantly associated with increased odds. However, since the estimated coefficient of alcoholic intake is not statistically significant ($p = 0.0872 > 0.05$), we cannot conclude that the alcoholic usage would cause an effect on the presence of cardiovascular diseases. In conclusion, the new model without suspicious high influential points and outliers can also produce the same results and interpretations. The final selected logistic regression model can be used to test if the life habits are significantly associated with the cardiovascular diseases.
In addition to the previous logistic regression model, we can also discrete the BMI covariates to select and

then get the following final model:

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{I[gender = MALE]} + \beta_3 \text{I[BMI='underweight']} + \beta_4 \text{I[BMI='overweight']} + \beta_5 \text{I[smoke=NON-SMOKE]} + \beta_6 \text{I[alco=YES]} + \beta_7 \text{I[active=YES]}$$

where it has least $BIC = 90524$. The new model fitted results can be shown in Table 4. The model diagnostics results (See Appendix-1) show that there is no significant indication of lack of fitting in the model with discrete BMI. Similar to the previous model, we can also draw the same conclusion that: being male, smoker, without regular physical activities and increased age can be significantly associated with the increased odds of the cardiovascular disease. However, since the estimated coefficient of alcoholic intake is not statistically significant ($p = 0.3196 > 0.05$), we cannot conclude that the alcoholic usage would cause an effect on the presence of cardiovascular diseases. It shows that the 'underweight' BMI has negative effect of the log odds and the 'overweight' BMI tends to have larger odds compared with the normal BMI when fixing other covariates.

Table 4: Model Fitting Results of Logistic Regression Model with Discrete BMI

-	Estimate β	Standard Error	z-statistic	p-value
(Intercept)	-4.014950	0.067684	-59.319	< 0.001
age	0.070886	0.001213	58.434	< 0.001
gender (=MALE)	0.106004	0.017700	5.989	< 0.001
BMI='underweight'	-0.485668	0.091775	-5.292	< 0.001
BMI='overweight'	0.600273	0.016677	35.994	< 0.001
smoke (=NON-SMOKER)	-0.075907	0.031163	-2.436	0.0149
alco (=YES)	-0.037162	0.037338	-0.995	0.3196
active (=YES)	-0.178659	0.019975	-8.944	< 0.001

In summary, the model comparison shows that we can draw consistent conclusion for the association of life habits and the cardiovascular diseases.

4 Discussion

4.1 Model Interpretation and Conclusion

From the previous modeling methods, the selected logistic model with least BIC can be used to analyze the relationships between life habits and the presence of cardiovascular diseases. The conclusions made by the models are robust and consistent between different models. The analysis of survey data including 69269 records about cardiovascular diseases has revealed some factors that associated with the presence of cardiovascular disease including being male, advanced age, being smoker, without regular physical activities and high BMI. This conclusion agree with some clinical studies with small sample size. [2] However, the intake of alcoholic tends to have no significant effect on the cardiovascular disease.

4.2 Limitation and Further Work

Even though the model shows no significant relationships between alcoholic intake and cardiovascular diseases, some clinical trial has the opposite conclusion. [3] This is because the alcoholic usage in our survey data is defined as alcoholism, compared with other studies that record the frequency or amount of alcohol consumption. The survey data are not conducted in randomized clinical trial, which means the association has not causal links between the predictors and response. In this case, we cannot conclude that the physical activities or smoking would lead to the cardiovascular diseases.
In this study, the interaction effect and subgroup analysis are waiting to be done in the future. Instead of fitting logistic regression model directly, the smoothing methods can be also applied in the continuous covariates to improve the goodness-of-fit. In this case, more studies are needed to generalize the results and make causal inference.

Reference

[1] Grundy, S. M., Pasternak, R., Greenland, P., Smith, S., Fuster, V. (1999). Assessment of cardiovascular risk by use of multiple-risk-factor assessment equations: a statement for healthcare professionals from the American Heart Association and the American College of Cardiology. Journal of the American College of Cardiology, 34(4), 1348-1359.

[2] Sundquist, J., Malmström, M., Johansson, S. E. (1999). Cardiovascular risk factors and the neighbourhood environment: a multilevel analysis. International journal of epidemiology, 28(5), 841-845.

[3] Ronksley, P. E., Brien, S. E., Turner, B. J., Mukamal, K. J., Ghali, W. A. (2011). Association of alcohol consumption with selected cardiovascular disease outcomes: a systematic review and meta-analysis. Bmj, 342, d671.

5 Appendix-1: Diagnostics of model with discrete BMI

```
        Runs Test - Two sided

data:  res.P
Standardized Runs Statistic = -1.5844, p-value = 0.1131


        Runs Test - Two sided

data:  res.D
Standardized Runs Statistic = -1.5844, p-value = 0.1131
```

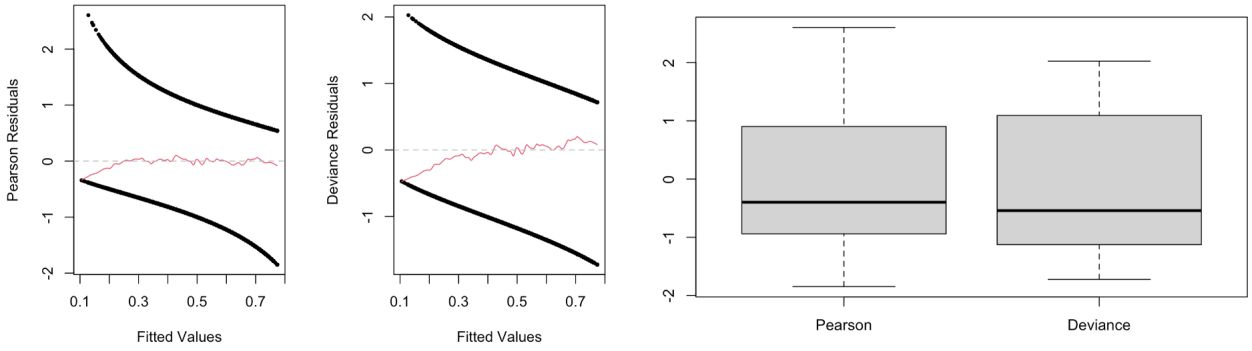

Figure 4: Model Diagnostics Results

6 Appendix-2: Diagnostics of model without quadratic term of BMI

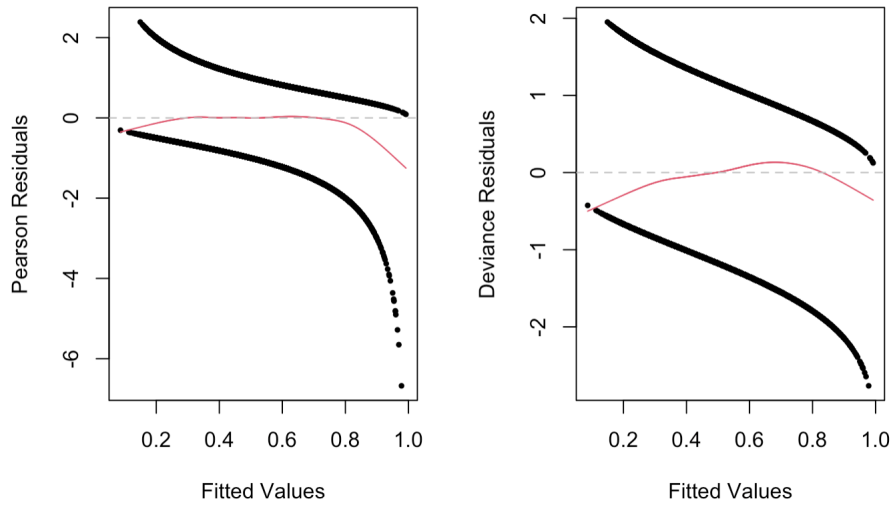

Figure 5: Model Diagnostics Results

7 Appendix-R-code

```
## Exploratory Data Analysis
```{r, echo = TRUE}
## IMPORT CVD DATASET
library(readr)
library(tidyverse)
library(broom)
library(MASS)
cardio_train <- read_delim("cardio_train.csv",
    ";", escape_double = FALSE, col_types = cols(gender = col_factor(levels = c("1",
        "2")), smoke = col_factor(levels = c("0",
        "1")), alco = col_factor(levels = c("0",
        "1")), active = col_factor(levels = c("0",
        "1")), cardio = col_factor(levels = c("0",
        "1"))), trim_ws = TRUE)
DATA <- na.omit(cardio_train)
DATA <- DATA[,c(2,3,4,5,10,11,12,13)]
```
```

```
```{r, echo = TRUE}
## Summary Statistic for Each Features
DATA['BMI'] <- DATA$weight/(DATA$height/100)^2
DATA$age <- as.integer(DATA$age/365)
DATA <- DATA[DATA$BMI>8 & DATA$BMI<105,] # delete the extreme value of impossible BMI
DATA <- DATA[,-c(3,4)]
summary(DATA)
hist(DATA$age,breaks=5,freq = FALSE)
```
Histogram
```{r}
# age and CVD
DATA_age_cardio <- as.data.frame(DATA$age[DATA$cardio==1])
colnames(DATA_age_cardio) <- 'age'
## Age and CVD: old people tend to have CVD
ggplot(data=DATA_age_cardio, mapping = aes(x=age))+
  geom_histogram(bins = 7)+
  labs(x ="Age in years", y = "CVD Status")+
  theme_linedraw()

# age and BMI

DATA_BMI_cardio <- as.data.frame(DATA$BMI[DATA$cardio==1])
colnames(DATA_BMI_cardio) <- 'BMI'
## Age and CVD: old people tend to have CVD
ggplot(data=DATA_BMI_cardio, mapping = aes(x=BMI))+
  geom_histogram()+
  labs(x ="BMI", y = "CVD Status")+
  theme_linedraw()
for (i in 1:dim(DATA)[1]) {
  if(DATA$BMI[i]<18.5){
    DATA$BMI_DISCRETE[i] <- 2
  } else if(DATA$BMI[i]< 24.9) {
    DATA$BMI_DISCRETE[i] <- 1
  } else{
    DATA$BMI_DISCRETE[i] <- 3
  }
}
DATA$BMI_DISCRETE <- as.factor(DATA$BMI_DISCRETE )
```
```

Exploratory Data Analysis

```{r, echo = TRUE}
## IMPORT CVD DATASET
library(readr)
library(tidyverse)
library(broom)
library(MASS)
cardio_train <- read_delim("cardio_train.csv",
    ";", escape_double = FALSE, col_types = cols(gender = col_factor(levels = c("1",
        "2")), smoke = col_factor(levels = c("0",
        "1")), alco = col_factor(levels = c("0",
        "1")), active = col_factor(levels = c("0",
        "1")), cardio = col_factor(levels = c("0",
        "1"))), trim_ws = TRUE)
DATA <- na.omit(cardio_train)
DATA <- DATA[,c(2,3,4,5,10,11,12,13)]
```

```{r, echo = TRUE}
## Summary Statistic for Each Features
DATA['BMI'] <- DATA$weight/(DATA$height/100)^2
DATA$age <- as.integer(DATA$age/365)
DATA <- DATA[DATA$BMI>8 & DATA$BMI<105,] # delete the extreme value of impossible BMI
DATA <- DATA[,-c(3,4)]
summary(DATA)
hist(DATA$age,breaks=5,freq = FALSE)
```

Histogram

```{r}
# age and CVD
DATA_age_cardio <- as.data.frame(DATA$age[DATA$cardio==1])
colnames(DATA_age_cardio) <- 'age'
## Age and CVD: old people tend to have CVD
ggplot(data=DATA_age_cardio, mapping = aes(x=age))+
  geom_histogram(bins = 7)+
  labs(x ="Age in years", y = "CVD Status")+
  theme_linedraw()

# age and BMI

DATA_BMI_cardio <- as.data.frame(DATA$BMI[DATA$cardio==1])
colnames(DATA_BMI_cardio) <- 'BMI'
## Age and CVD: old people tend to have CVD
ggplot(data=DATA_BMI_cardio, mapping = aes(x=BMI))+
  geom_histogram()+
  labs(x ="BMI", y = "CVD Status")+
  theme_linedraw()
for (i in 1:dim(DATA)[1]) {
  if(DATA$BMI[i]<18.5){
    DATA$BMI_DISCRETE[i] <- 2
  } else if(DATA$BMI[i]< 24.9) {
    DATA$BMI_DISCRETE[i] <- 1
  } else{
    DATA$BMI_DISCRETE[i] <- 3
  }
}
DATA$BMI_DISCRETE <- as.factor(DATA$BMI_DISCRETE )
```

Logistic Regression

```{r, echo = TRUE}
# logistic regression with all first-order predictor
lr_model1 <- glm(cardio ~ age+gender+BMI+I(BMI^2)+smoke+alco+active, data = DATA, family = 'binomial')
summary(lr_model1)
# model selection methods with AIC
Scope = list(upper = ~ (age+gender+BMI+smoke+alco+active), lower = ~1)

# BIC
summary(lr_model_full)
lr_model_full <- glm(cardio ~ (age+gender+BMI+smoke+alco+active)^2+I(BMI^2)+I(age^2), data = DATA, family = 'binomial')
Scope_2 = list(upper = ~ (age+gender+BMI+smoke+alco+active)^2+I(BMI^2)+I(age^2), lower =
~(age+gender+BMI+smoke+alco+active))

lr_model_full <- glm(cardio ~ age + gender + BMI + active + age:gender + age:BMI + gender:BMI, data = DATA, family =
'binomial')
Scope_2 = list(upper = ~ age + gender + BMI + active + age:gender + age:BMI + gender:BMI, lower =
~(age+gender+BMI+smoke+alco+active))

BIC_models <- stepAIC(lr_model_full, trace = FALSE, scope = Scope_2, k=log(dim(DATA)[1]))
BIC_models$anova

lr_model2 <- glm(cardio ~ age+gender+BMI+I(BMI^2)+smoke+active+alco, data = DATA, family = 'binomial')
summary(lr_model2)
```

```{r}
lr_model_s <- glm(cardio ~ age + gender + BMI_DISCRETE + smoke + alco + active, data = DATA, family = 'binomial')
summary(lr_model_s)
```

Logistic

```{r}
library(splines)
lr_model_s <- glm(cardio~age + gender + BMI +I(BMI^2) +smoke + alco + active,data=DATA,family = 'binomial')
summary(lr_model_s)
exp(lr_model_s$coefficients)
```

Logistic Diagnostics

```{r}
# We expect these two types of residuals have similar distributions.
# no lack-of-fit => similar boxplots
# similar boxplots -> next step: Residual Plots
model_diag <- lr_model_s
res.P = residuals(model_diag, type="pearson")
res.D = residuals(model_diag, type="deviance") #or residuals(fit), by default
boxplot(cbind(res.P, res.D), names = c("Pearson", "Deviance"))

# * Residual Plots ----------------
# no lack-of-fit => no systematic pattern
par(mfrow=c(1,2))
plot(model_diag$fitted.values, res.P, pch=16, cex=0.6, ylab='Pearson Residuals', xlab='Fitted Values')
lines(smooth.spline(model_diag$fitted.values, res.P, spar=1), col=2)
abline(h=0, lty=2, col='grey')
plot(model_diag$fitted.values, res.D, pch=16, cex=0.6, ylab='Deviance Residuals', xlab='Fitted Values')
lines(smooth.spline(model_diag$fitted.values, res.D, spar=1.1), col=2)
abline(h=0, lty=2, col='grey')

# * Runs Test ----------------

# Null hypothesis: no systematic pattern
# Reject H0 => lack-of-fit
# In the plot: consecutive positives/negatives => systematic pattern

library(lawstat)
# please pay attention to the library
# there are different runs.test() functions in different packages
# we are specifically using this one!

runs.test(y = res.P, plot.it = FALSE)
title(main='Pearson Residual Runs Test')
runs.test(y = res.D, plot.it = FALSE)
title(main='Deviance Residual Runs Test')


```
```{r}
# leverage points => influential points

leverage = hatvalues(model_diag)
plot(names(leverage), leverage, xlab="Index", type="h")
points(names(leverage), leverage, pch=16, cex=0.6)
p <- length(coef(model_diag))
n <- nrow(DATA)
abline(h=2*p/n,col=2,lwd=2,lty=2)
infPts <- which(leverage>2*p/n)
```

```{r}
# ** Cook's Distance ----------------

# high Cook's distance => influential points/outliers
# leverage points with high Cook's distance => suspicious influential points & outliers
#                     may need to be deleted -> check scatterplots

cooks = cooks.distance(model_diag)

plot(cooks, ylab="Cook's Distance", pch=16, cex=0.6)
points(infPts, cooks[infPts], pch=17, cex=0.8, col=2)
susPts <- as.numeric(names(sort(cooks[infPts], decreasing=TRUE))[1:3])
text(susPts, cooks[susPts], susPts, adj=c(-0.1,-0.1), cex=0.7, col=4)

```
```{r}
DATA[susPts,]
hist(DATA[infPts,]$BMI)

lr_model_s <- glm(cardio ~ age+gender+BMI+I(BMI^2)+smoke+alco+active, data = DATA, family = 'binomial')
summary(lr_model_s)

DATA_3 <- DATA[-susPts,]
lr_model_s_3 <- glm(cardio ~ age+gender+BMI+I(BMI^2)+smoke+alco+active, data = DATA_3, family = 'binomial')
summary(lr_model_s_3)

DATA_all <- DATA[-infPts,]
summary(DATA_all)
table(DATA_all$smoke,DATA_all$cardio)
lr_model_s_all <- glm(cardio ~ age+gender+BMI+I(BMI^2)+smoke+active, data = DATA_all, family = 'binomial')
summary(lr_model_s_all)

```