Understanding the association between the risk factors of Cardiovascular diseases in the United States

HDS 5310-04

Professor in charge: Paul Boal

Saint Louis university School of Medicine

INTRODUCTION:

Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year[1]. In 2021, it resulted in the deaths of around 0.695 million individuals, constituting one-fifth of all deaths. The economic impact of heart diseases on the nation totals approximately $239.9 billion annually, encompassing healthcare expenses, medication costs, and productivity losses due to fatalities[2]. Elderly individuals aged 65 and above are significantly more prone to experiencing heart attacks, strokes, coronary heart disease, often referred to as heart disease, and heart failure compared to younger age groups[3]. According to WHO, CVDs are a group of disorders of the heart and blood vessels. They include coronary heart disease, cerebrovascular disease, peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis and pulmonary embolism[4].

Behavioral risk factors such as smoking, excessive alcohol consumption, poor dietary habits, lack of physical activity, and sedentary lifestyle, along with physiological risk factors like hypertension, diabetes, obesity, and high cholesterol, as well as non-modifiable factors such as age, gender, family history, and ethnicity, are each linked independently to CVD outcomes[5]. The prevalence of cardiovascular disease in women varies based on racial and ethnic backgrounds[6]. Global Evidence indicates that addressing and managing modifiable risk factors can potentially lead to a reduction of up to 90% in cases of cardiovascular diseases[6].

Among CVD risk factors, diabetes mellitus, hypertension, smoking and hyperlipidemia have emerged as the most significant public health challenges, exacerbating the risk for CVD complications [4, 7]. To mitigate the risk of heart disease, this

research focuses on identifying risk factors and the presence of heart disease, with emphasis on gender, age, blood pressure, cholesterol levels, and diabetes.

Understanding the relationship between development of CVDs associated with specific risk factors, such as gender disparities and age-related vulnerabilities, is essential for targeted prevention and intervention strategies to address the issue of CVD. There is an unclear opinion on the progression of CVDs based on few risk factors like hyperlipidemia, blood pressure, smoking and diabetes[7] or solely based on sex and age [8]. Hence, by focusing on all the major risk factors together like age, sex, diabetes, hypertension, and hyperlipidemia, the research aims to identify key areas for intervention and to improve cardiovascular health outcomes. Moreover, the acknowledgment of gaps in current knowledge underscores the need for further research to better elucidate the relationships between specific risk factors and CVD development, thereby guiding evidence-based approaches to mitigate the threat of heart diseases.

Based on the relevance and significance in earlier research on cardiovascular diseases (CVD), these independent variables were chosen. The prevalence, risk factors, and trends related to CVD have all been studied in the past; however, this study contributes to the existing literature by focusing on the development of heart disease based on the above-mentioned specific variables. The research question is important because it could help find groups of people who are vulnerable to developing heart diseases due to the specified possible risk factors and aid in developing targeted policies and interventions meant to reduce the deaths from CVDs. By studying how these risk factors are connected to CVDs, healthcare professionals and public health officials can develop improved strategies to treat and avoid the issues that are more effective for

vulnerable groups. The study's practical implications include the possibility of improving public health outcomes by modifying healthcare policies, laws, and individual behavioral change programs meant to lower the prevalence of cardiovascular diseases.

METHODOLOGY:

The main purpose of this research is to investigate the relationship between certain risk factors and the presence of heart disease, particularly focusing on Sex, Age, Blood Pressure, Cholesterol levels and Diabetes. These variables are chosen because these are considered major risk factors contributing to the development of heart disease based on the previous literature. The dependent variable in this research is heart disease, a categorical variable with two responses (presence/absence) and the independent variables are the gender which categorizes into male and female, age, blood pressure levels, cholesterol levels, presence, or absence of diabetes (fasting blood sugar over 120). To confirm the relationship between the outcome and predictor variables, statistical tests such as Chi-square tests and independent sample t-tests will be performed.

Descriptive statistics: The descriptive statistics of the entire data will be analyzed before and after recoding the appropriate variable categories by using the summary function. Prop. Table function will be used as a descriptive statistic for both the categorical variables (sex, and blood sugar/diabetes) that are to be included in the chi-square test of independence. To examine the relationship between two categorical variables, bar plots will be plotted showing the percentage of the category (sex, and blood sugar) with either having heart disease or not.

Data management: the file is imported by using the read csv function as the file is a csv one. Based on the descriptive statistics, the re-coding of the variables will be done

assigning the category names to the categorical variables according to the data dictionary so that the unnecessary values will be replaced with the categorical values. The transformations if needed will be performed on numeric variables such as age, blood pressure and cholesterol. The linear transformations such as log, square root, cube root, and inverse will all be performed until the normality distribution is achieved.

Implementation of statistical tests: The chi-square tests will be performed individually for the variables like sex, and diabetes with the outcome variable heart disease because both the variables in these tests are categorical with just two responses. The alternate and null hypothesis for these two tests are:

A. Hypotheses for the Chi-square test between Heart disease and sex

   H0: There is no relationship between the presence of heart disease and sex.

   HA: There is a relationship between the presence of heart disease and sex.

B. Hypotheses for the Chi-square test between Heart disease and diabetes

   H0: Presence of heart disease is the same among the groups with blood sugar levels over and less than 120.

   HA: Presence of heart disease is not the same among the groups with blood sugar levels over and less than 120.

After performing the chi-square test, the assumptions of those tests will be checked. The assumptions include:

- The variables must be nominal or ordinal
- The expected values should be 5 or higher in at least 80% of groups

- The observations must be independent.

The standardized residuals will also be calculated for both these variables using the Crosstable (contingency table) function which will indicate any deviations of expected and observed values. Then, the effect size will be calculated using Cramer'sV. The effect size may range from small to medium to large depending on the Cramer'sV value indicating the weak to moderate to strong relationship between the variables.

- Small or weak effect size for V = .1
- Medium or moderate effect size for V = .3
- Large or strong effect size for V = .5

Then, independent sample t-tests will be performed individually for the predictor variables like age, blood pressure and cholesterol levels with the outcome variable heart disease as one among each test variable are continuous and the outcome variable is commonly categorical in all these three cases. The alternate and null hypotheses for these three tests are:

A. Hypotheses for the independent sample t-test between the Heart disease and age

   H0: There is no difference in the mean age between presence and absence of heart disease.

   HA: There is a difference in the mean age between presence and absence of heart disease.

B. Hypotheses for the independent sample t-test between the Heart disease and blood pressure.

H0: There is no difference in the mean blood pressure level between presence and absence of heart disease.

HA: There is a difference in the mean blood pressure level between presence and absence of heart disease.

C. Hypotheses for the independent sample t-test between the Heart disease and cholesterol.

H0: There is no difference in the mean cholesterol level between presence and absence of heart disease.

HA: There is a difference in the mean cholesterol level between presence and absence of heart disease.

The assumptions for the independent sample t-tests are:

- Continuous variable and two independent groups
- Independent observations
- Normal distribution in each group
- Equal variances for each group.

If any variable fails one assumption, an alternate test like Wilcoxon rank sum test will be performed. Then, the effect size will be measured using Cohen's d. Its values indicate the effect size as follows:

- 2 to d < .5 is a small effect size
- .5 to d < .8 is a medium effect size

- ≥ .8 is a large effect size.

The hypotheses for such alternate tests will be framed and the significance is checked with the help of p-value. Then, the effect size is calculated for this alternate test using qnorm. Cramer's V will measure association strength for heart disease-gender and heart disease-diabetes. Cohen's d will assess the practical significance of differences in age, blood pressure, and cholesterol levels between heart disease groups. Effect size interpretations will follow established guidelines to determine relationship strength accurately.

Data visualizations: Visualizations such as histograms will be used to understand the descriptive statistics of the continuous variables (age, blood pressure and cholesterol) with the outcome variable, box plot will also be plotted using the continuous variables to see the distribution of age range by blood pressure category and Heart disease status as this can reveal any differences in age and blood pressure distributions between the two groups, providing insights into the relationship between these variables and heart disease risk, a violin plot will also be plotted to see the distribution of cholesterol by blood pressure category and Heart disease status as this can show any differences in cholesterol distributions among individuals with varying blood pressure levels and heart disease status, aiding in identifying potential risk factors associated with heart disease development and bar plots will also be used to see the distribution of sex, having fasting blood sugar over 120 in patients with heart disease and without the heart disease to identify potential gender disparities and the impact of diabetes on heart disease risk .

Final model: The model for this study will be a logistic regression model as the outcome variable is categorical. And the model significance and model fit representing the odds ratio and 97.5% and 2.5% confidence intervals will also be interpreted. The assumptions for logistic regression are:

- Independence of observations
- Linearity, and
- No perfect multicollinearity

As this study has categorical outcome variables, it is not possible to check the linearity. The model with all the significant variables will be compared with another model having only the significant odds ratio and efficiency of these models will be obtained by running a likelihood ratio test. Data frames will be created to check the models.

RESULTS:

GENDER

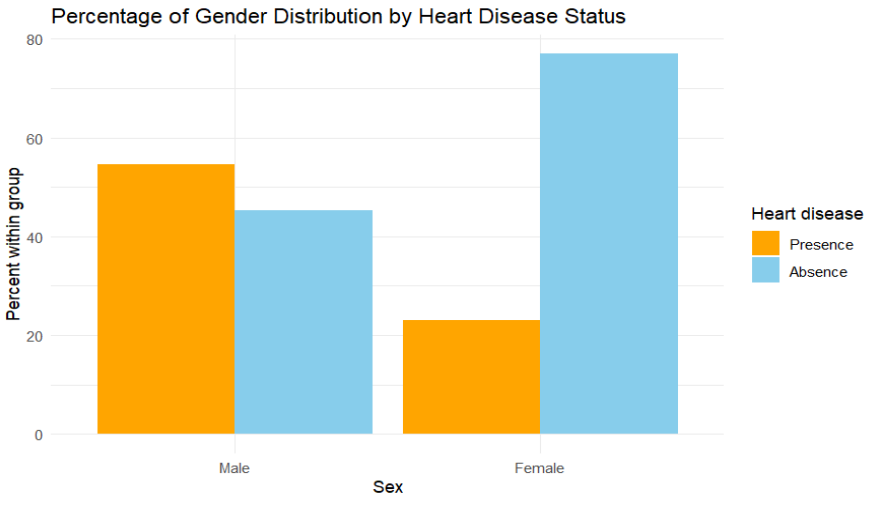

Fig1: Percentage of gender distribution by heart disease status

This study examined the association between heart disease and gender using a Chi-squared test of independence. Results revealed a significant difference in the presence of heart disease between males and females (X-squared = 22.667, $p < 0.001$). Fig 1 showed a higher percentage of males, and a lower percentage of females were having heart disease, indicating a significant association between gender and heart disease prevalence. Additionally, the assumptions for the chi-square test were also met. Specifically, the standardized residuals showed that the observed count of males with heart disease was significantly higher than expected (std. residual = 2.070), while for females, it was significantly lower than expected (std. residual = -3.002). Pearson's chi-squared test reaffirmed this association ($p < 0.05$). Additionally, Cramer's V, a measure of association for categorical variables, was approximately 0.29, indicating a moderate effect size and suggesting a meaningful relationship between heart disease and gender.

AGE

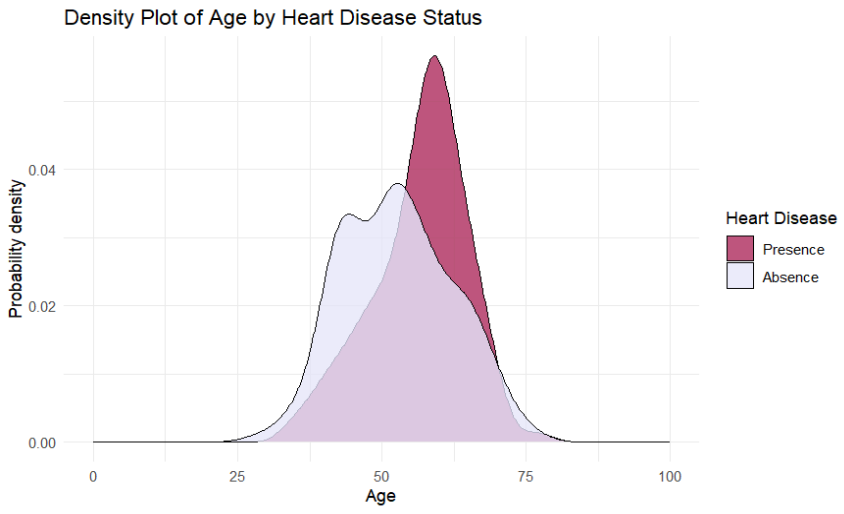

Fig 2: Density plot of age by heart disease status

This study also investigated the difference in mean age between individuals with and without heart disease in the United States. Fig 2 indicates that the patients with heart disease tend to be in the age range 50-70 on average than those without heart disease but it also indicates that the risk of having heart disease starts at the age of 30. The Welch Two Sample t-test indicated a significant difference in mean age between individuals with and without heart disease (t = 3.6199, df = 266.86, p = 0.0003526). The mean age for those with heart disease was approximately 56.59 years, compared to 52.71 years for those without. This result led to the rejection of the null hypothesis, suggesting a statistically significant difference in mean age. The effect size, calculated using Cohen's d, was approximately 0.44, indicating a moderate effect. The assumptions for t-test such as normality and equal variances were not met so, the alternate test which is Wilcoxon rank sum test yielded (W= 11366, p-value = 0.0002052) statistical significance between age and the outcome variable. The effect size for this test is 0.226 indicating a moderate relationship between heart disease and age.

DIABETES/ FASTING BLOOD SUGAR OVER 120

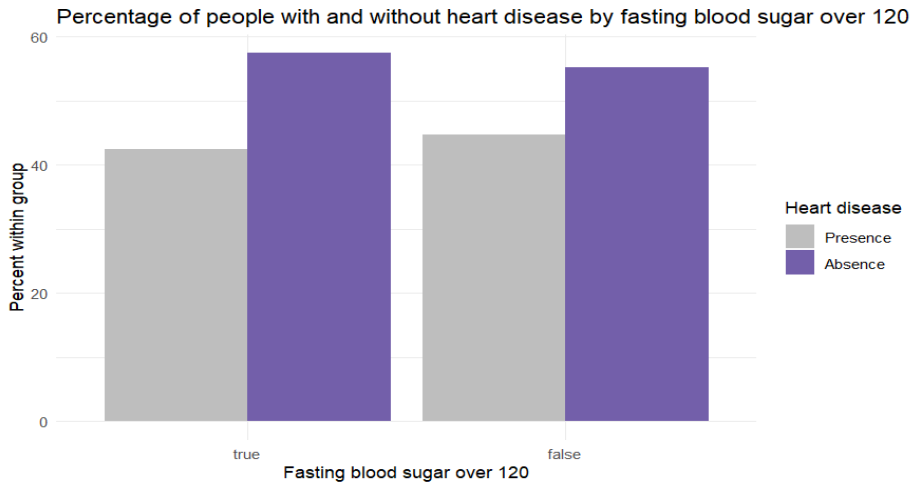

Fig 3: Percentage of people with and without heart disease by fasting blood sugar over 120

This study also investigated the association between the presence of heart disease and fasting blood sugar levels categorized as over or under 120 mg/dl using a chi-squared test of independence. Results revealed there is no statistically significant difference in the presence of heart disease between diabetic and non-diabetic patients (X-squared = 0.0091712, $p > 0.05$). Fig 3 showed that those with fasting blood sugar over 120 mg/dL, a higher percentage has the absence of heart disease compared to the presence of heart disease. Since there is no relationship, the effect size (0.0058) also found to be weak. But, the assumptions were all met. The standardized residuals also turned out to be indicating that there is no significant relationship between heart disease and fasting blood sugar over 120 (diabetes).

BLOOD PRESSURE

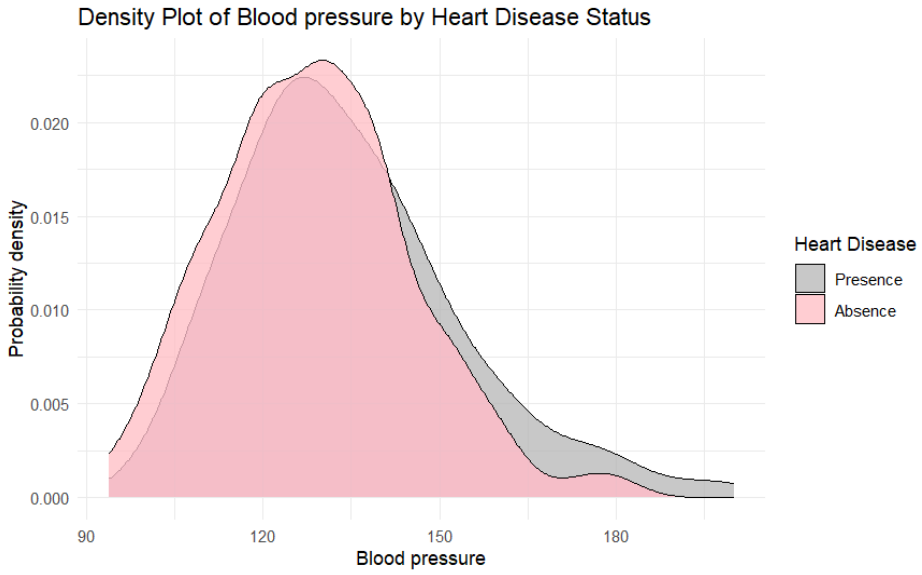

Fig 4: Distribution of cholesterol by blood pressure category and heart disease status

This study also investigated the association between the presence of heart disease and blood pressure (continuous variable) using an independent sample t-test. Fig 4 explains the association between both cholesterol and blood pressure with heart disease. The Welch Two Sample t-test indicated a significant difference in mean blood pressure level between individuals with and without heart disease ($t = 2.533$, $df = 235.92$, $p < 0.05$). The mean blood pressure for those with heart disease was approximately 134.44 mmHg, compared to 128.87mmHg for those without. This result led to the rejection of the null hypothesis, suggesting a statistically significant difference in mean blood pressure. The effect size, calculated using Cohen's d, was approximately 0.31, indicating a weak relationship. The assumptions for t-test were met except the normality so, the alternate test which is Wilcoxon rank sum test yielded ($W= 10368$, p-value = 0.031) statistical significance between blood pressure and the outcome variable. The effect size for this test is 0.131 indicating a weak relationship between heart disease and blood pressure.

CHOLESTEROL

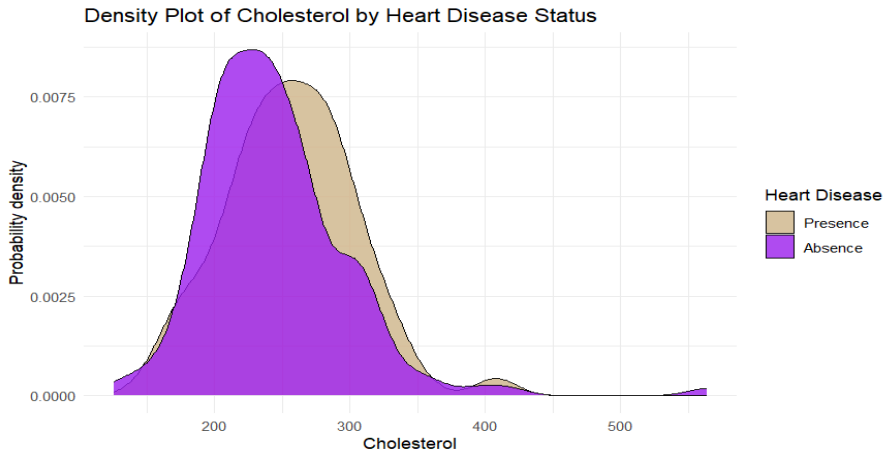

Fig 5: Distribution of cholesterol by age category and heart disease status

This study aimed to investigate the association between the presence of heart disease and cholesterol (continuous variable) using an independent sample t-test. Fig 5 explains the association between both cholesterol and age range with heart disease. The Welch Two Sample t-test indicated that there is no significant difference in mean cholesterol level between individuals with and without heart disease (t = 1.9715, df = 265.06, p-value = 0.05) because the p-value is almost 0.05. This result led to the retention of the null hypothesis, suggesting no statistically significant difference in mean cholesterol. The effect size, calculated using Cohen's d, was approximately 0.24, indicating a weak relationship. The assumptions for t-test were met except the normality so, the alternate test which is Wilcoxon rank sum test yielded (W= 10700, p-value = 0.007701) statistical significance between cholesterol and the outcome variable. The effect size for this test is 0.162 indicating a weak relationship between heart disease and blood pressure.

Final model-Logistic regression model:

The descriptive statistics revealed that the cholesterol and fasting blood sugar over 120 variables are not significant (p-value: >0.05) so, a model is built using the remaining variables that are age, sex, and blood pressure with the outcome variable. This model is considered a large model and it got the Count R-squared (percent correctly predicted): 68.89%, model sensitivity: 62.5%, and Model specificity: 74%. Age (OR= 1.058; 95% CI: 1.026-1.093) has significant odds ratio greater than 1, Sex (OR = 0.184; 95% CI: 0.095-0.339) has significant odds ratio less than 1, and Blood pressure (OR= 1.0165; 95% CI: 1.001-1.032). The assumptions like independence of observations, and multicollinearity were met for all the three variables except the linearity assumption for

age and blood pressure. Another small model was built including only sex variable with the outcome variable whose Count R-squared (percent correctly predicted) was 61.85%, model sensitivity was 83.4%, and Model specificity was 44.67%. According to this model, females have approximately 75% lesser odds of heart disease compared to males (OR = 0.247; 95% CI: 0.136-0.435). This variable has a significant odds ratio less than 1. When the likelihood ratio test was performed between these two models, the test statistic of $\chi 2 = 22.551$ had a p-value of <0.05 suggesting the larger model is better at predicting heart disease compared to the small model. Example data frames are created to check both the models to predict the percentage of having heart disease. Larger model considers age, sex, and blood pressure variables in predicting the presence of heart disease. According to this model, a female aged 60-years with 183 mm Hg blood pressure has 18% more risk compared to a male aged 20-years with 169 mm Hg blood pressure. Smaller model considers only sex variable in predicting the presence of heart disease. According to this model, males have 32% more probability/risk of having heart disease.

DISCUSSION:

The findings of this study contribute to the existing literature by highlighting the importance of risk factors in developing cardiovascular/heart disease among a cohort of patients revealing the significant relationship between the sex, age, and blood pressure. This study revealed a significant difference in the presence of heart disease between males and females, with males showing a higher percentage of heart disease prevalence. This finding underscores the importance of considering gender disparities in cardiovascular health outcomes and the need for targeted interventions to address this disparity. Furthermore, the study found a significant difference in mean age between

individuals with and without heart disease. The average age of heart disease patients was found to be higher, indicating that age is a major risk factor for the development of heart disease. This study highlighted the higher blood pressure was associated with an increased likelihood of heart disease, emphasizing the importance of hypertension management and blood pressure control in preventing cardiovascular complications. However, while cholesterol is a well-established risk factor for heart disease, the lack of significance in this study may indicate the need for further investigation or consideration of additional factors influencing cholesterol levels. Similarly, the study did not find a significant difference in the presence of heart disease based on fasting blood sugar levels categorized as over 120 mg/dl. This suggests that while diabetes is a known risk factor for heart disease, other factors may play a more prominent role in predicting heart disease prevalence in this population. The logistic regression model further supported the significance of age, gender, and blood pressure as predictors of heart disease. The larger model including age, sex, and blood pressure performed better in predicting heart disease compared to the smaller model containing only sex as a predictor. This highlights the importance of considering multiple risk factors simultaneously for accurate risk assessment and prediction of heart disease.

CONCLUSION:

In conclusion, this study provides valuable insights into the relationship between various risk factors and the presence of heart disease. The prevalence of heart disease was found to be significantly predicted by blood pressure, gender, and age. This highlights the significance of preventative methods and focused therapies that are specific to particular demographic groups. Improved outcomes, such as higher survival rates and a greater

chance of effective treatments, are linked to early detection and treatment of cardiac disease. So, the population at higher risk should be properly screened from time to time for early detection of heart disease. While the study contributes to our understanding of cardiovascular risk factors, more investigation is required to explore other variables, such as lifestyle choices, genetic predisposition, and socioeconomic circumstances, that may influence the development of heart disease. Healthcare practitioners and policymakers can better reduce the burden of cardiovascular diseases and enhance public health outcomes by identifying the multifaceted causes of heart disease and addressing them.

REFERENCES

1. World Health Organization. Cardiovascular Diseases. World Health Organization. Published 2023. https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
2. Centers for Disease Control and Prevention. Heart Disease Facts. Centers for Disease Control and Prevention. Published May 15, 2023. https://www.cdc.gov/heartdisease/facts.htm
3. Alyami SS, Algharbi A, Alsuwaidan S. Characteristics of Associated Diseases in Older Patients with Cardiovascular Disease. *Advances in Aging Research*. 2022;11(6):151-161. doi:https://doi.org/10.4236/aar.2022.116011
4. World Health Organization. Cardiovascular diseases (CVDs). World Health Organization. Published June 11, 2021. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
5. Khalid Abdul Basit, Linda Ng Fat, Gregg EW. Changes in cardiovascular risk factors for diabetes among young versus older English adult populations. *Zeitschrift für Gesundheitswissenschaften/Journal of public health*. Published online December 23, 2023. doi:https://doi.org/10.1007/s10389-023-02143-5
6. Abbas Rezaianzadeh, Moftakhar L, Seif M, Masoumeh Ghoddusi Johari, Seyed Vahid Hosseini, Seyed Sina Dehghani. Incidence and risk factors of cardiovascular disease among population aged 40–70 years: a population-based cohort study in the South of Iran. *Tropical Medicine and Health*. 2023;51(1). doi:https://doi.org/10.1186/s41182-023-00527-7

7. Wei M, Mitchell BD, Haffner SM, Stem MP. Effects of Cigarette Smoking, Diabetes, High Cholesterol, and Hypertension on All-Cause Mortality and Cardiovascular Disease Mortality in Mexican AmericansThe San Antonio Heart Study. *American Journal of Epidemiology*.1996;144(11):1058-1065. doi:https://doi.org/10.1093/oxfordjournals.aje.a008878
8. Jousilahti, P. *et al.* (1999) 'Sex, age, cardiovascular risk factors, and coronary heart disease', *Circulation*, 99(9), pp. 1165–1172. doi:10.1161/01.cir.99.9.1165.

DATA:

Heart disease prediction - dataset by Informatics-Edu. data.world. August 24, 2020. Accessed May 6, 2024. https://data.world/informatics-edu/heart-disease-prediction.

DATA DICTIONARY:

Diabetes Prediction - dataset by informatics edu | data.world. data.world/informatics-edu/diabetes-prediction. Accessed May 6, 2024.

APPENDIX

```
---
title: "Final Group project"
author: "Venkata Sai Pallavi Pallapolu, Supraja Medicherla, Sai Teja Yadav Gajji"
date: "2024-04-30"
output: html_notebook
---

# Prologue

-   **PROJECT:** final group research project
-   **PURPOSE:** building a model using the significant variables for predicting the heart disease
-   **DATA:** Heart_Disease_prediction.csv
-   **AUTHOR:** Venkata Sai Pallavi Pallapolu, Supraja Medicherla, Sai Teja Yadav Gajji
-   **CREATED:** 2024-04-30
---
```

```
## Loading the necessary packages

```{r}
#install the necessary packages
packages <- c("tidyverse", "car", "dunn.test", "tableone", "dplyr", "Hmisc", "descr","rcompanion")
```

```{r}
purrr::walk(packages, library, character.only = T)
```

```{r}
library("tidyverse")
library(readr)
library("reshape2")
library(ggplot2)
library(car)
library("dplyr")
library("odds.n.ends")
```
```

```
**Interpretation:** This data set is used to predict heart disease.
Patients were classified as having or not having heart disease based on cardiac catheterization, the gold standard.
If they had more than 50% narrowing of a coronary artery they were labeled as having heart disease.

-   The data set can be accessed through this link:
<https://data.world/informatics-edu/heart-disease-prediction/workspace/file?filename=Heart_Disease_Prediction.csv>. It also provides the data dictionary that explains about each variable.

## Descriptive statistics for the data before recoding

```{r}
# Reading the data file
heart <- read.csv("Heart_Disease_Prediction.csv")
summary(heart)
```
```

```
## Recoding the data using the data dictionary

```{r}
#Re-coding
heart.cleaned <- heart %>%
  mutate(Sex = recode_factor(.x = Sex, `1` = "Male",
                                       `0` = "Female")) %>%
  mutate(Chest.pain.type = recode_factor(.x = Chest.pain.type,`1` = "typical angina",
                                         `2` = "atypical angina",
                                         `3` = "non-anginal pain",
                                         `4` = "asymptomatic")) %>%
  mutate(FBS.over.120 = recode_factor(.x = FBS.over.120,`1` = "true",
                                      `0` = "false")) %>%
  mutate(EKG.results = recode_factor(.x = EKG.results, `0` = "normal",
                                     `1` = "having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)",
                                     `2` = "showing probable or definite left ventricular hypertrophy")) %>%
  mutate(Exercise.angina = recode_factor(.x = Exercise.angina, `1` = "yes",
                                         `0` = "no")) %>%
  mutate(Slope.of.ST = recode_factor(.x = Slope.of.ST, `1` = "upsloping",
                                     `2` = "flat",
                                     `3` = "downsloping")) %>%
  mutate(Thallium = recode_factor(.x = Thallium, `3` = "normal",
                                  `6` = "fixed defect",
                                  `7` = "reversable defect")) %>%
  mutate(Heart.Disease = recode_factor(.x = Heart.Disease,`0` = "< 50% diameter narrowing",
                                       `1` = "> 50% diameter narrowing"))
heart.cleaned
```
```

```
## Checking the descriptive statistics after recoding

```{r}
summary(heart.cleaned)
```
```

```
## Using graphs to examine the relationship between outcome and predictor variables

```{r}
#Barplot to show the percent of gender of individuals with and without heart disease
heart.sex.plot <- heart.cleaned %>%
drop_na(Heart.Disease) %>%
drop_na(Sex) %>%
group_by(Heart.Disease, Sex) %>%
count() %>%
group_by(Sex) %>%
mutate(perc = 100*n/sum(n)) %>%
ggplot(aes(x = Sex, y = perc, fill = Heart.Disease)) +
geom_bar(position = "dodge", stat = "identity") +
theme_minimal() +
scale_fill_manual(values = c("orange", "skyblue"),
                  name = "Heart disease") +
labs(x = "Sex", y = "Percent within group",title = "Percentage of Gender Distribution by Heart Disease Status")
heart.sex.plot
```

```{r}
#Density Plot of Age by Heart Disease Status
heart.cholesterol.plot <- heart.cleaned %>%
  ggplot(aes(x = Cholesterol,
             fill = Heart.Disease)) +
  geom_density(alpha = .8) +
  theme_minimal() +
  labs(x = "Cholesterol", y = "Probability density", fill= "Heart disease", title = "Density Plot of Cholesterol by Heart Disease Status") +
  scale_fill_manual(values = c('tan', 'purple'),
                    name = "Heart Disease")
heart.cholesterol.plot
```

```{r}
#Barplot to show the percent of individuals with and without heart disease having fasting blood sugar over 120
heart.diabetes.plot <- heart.cleaned %>%
drop_na(Heart.Disease) %>%
drop_na(FBS.over.120) %>%
group_by(Heart.Disease, FBS.over.120) %>%
count() %>%
group_by(FBS.over.120) %>%
mutate(perc = 100*n/sum(n)) %>%
ggplot(aes(x = FBS.over.120, y = perc, fill = Heart.Disease)) +
geom_bar(position = "dodge", stat = "identity") +
theme_minimal() +
scale_fill_manual(values = c("gray", "#7463AC"),
                  name = "Heart disease") +
labs(x = "Fasting blood sugar over 120", y = "Percent within group", title = "Percentage of people with and without heart disease by fasting blood sugar over 120")
heart.diabetes.plot
```

```{r}
#Density Plot of Age by Heart Disease Status
heart.bp.plot <- heart.cleaned %>%
  ggplot(aes(x = BP,
             fill = Heart.Disease)) +
  geom_density(alpha = .8) +
  theme_minimal() +
  labs(x = "Blood pressure", y = "Probability density", fill= "Heart disease", title = "Density Plot of Blood pressure by Heart Disease Status") +
  scale_fill_manual(values = c('grey', 'pink'),
                    name = "Heart Disease")
heart.bp.plot
```

Statistical significant tests for the outcome and predictor variables

1) Chi-squared test of independence

Descriptive statistics for two categorical variables

```{r}
#Descriptive statistics
table(heart.disease=heart.cleaned$Heart.Disease,gender=heart.cleaned$Sex)
```

H0: There is no relationship between presence of heart disease and sex

HA: There is a relationship between presence of heart disease and sex

```{r}
#Chi-squared test of independence for the variables Heart disease and Sex
chisq.test(x=heart.cleaned$Heart.Disease, y=heart.cleaned$Sex)
```

```
#Chi-squared examining the presence of heart disease and sex using standardized residuals
CrossTable(heart.cleaned$Heart.Disease, heart.cleaned$Sex, expected = TRUE, prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE, chisq = TRUE, sresid = TRUE)
```
```

```r
#Calculating the effect size using cramer's V
library("lsr")
cramersV(x= heart.cleaned$Heart.Disease,
         y= heart.cleaned$Sex)
```

```r
# descriptive stats using histogram
heart.age.plot <- heart.cleaned %>%
  ggplot(aes(x = Age,fill=Heart.Disease)) +
  geom_histogram(color = "white") +
  theme_minimal() +
  labs(x = "Age", y = "Patients")
heart.age.plot
```

```{r}
#mean of age for both presence of heart disease and no heart disease
mean_age_heart_disease <- mean(heart.cleaned$Age[heart.cleaned$Heart.Disease == "Presence"], na.rm = TRUE)
mean_age_no_heart_disease <- mean(heart.cleaned$Age[heart.cleaned$Heart.Disease == "Absence"], na.rm = TRUE)
print(mean_age_heart_disease)
print(mean_age_no_heart_disease)
```

H0: There is no difference in the mean age between the people with and without heart disease in the united states.

HA: There is a difference in the mean age between the people with and without heart disease in the united states.

```{r}
#Comparing the age for the presence and absence of heart disease using t-test
t.test(formula=heart.cleaned$Age~ heart.cleaned$Heart.Disease)
```

Source Visual Outline

```
hrt.age.plot <- heart.cleaned %>%
  ggplot(aes(x = Age,
             fill = Heart.Disease)) +
  geom_density(alpha = .8) +
  theme_minimal() +
  labs(x = "Age", y = "Probability density", fill= "Heart disease", title = "Density Plot of Age by Heart Disease Status") +
  scale_fill_manual(values = c('maroon', 'lavender'),
                    name = "Heart Disease") +
  xlim(0,100)
hrt.age.plot
```

```{r}
#Calculating effectsize using cohensD
lsr::cohensD(x=Age~Heart.Disease, data = heart.cleaned, method = "unequal")
```

Normal distribution

```{r}
#Checking the normality with the histogram
heart.cleaned %>%
  ggplot(aes(x=Age))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Age",
       y="Frequency")
```

Source Visual Outline

```
heart.cleaned %>%
  drop_na(Age) %>%
  ggplot(aes(sample = Age))+
  stat_qq(aes(color = "Frequency of Age groups", alpha = .6)) +
  facet_grid(cols = vars(Heart.Disease)) +
  geom_abline(aes(intercept = mean(x= Age),
                  slope = sd(x=Age), linetype= "Normally distributed"),
              color="gray", size=1)+
  theme_minimal()+
  labs(x="theoretical normal distribution", y="Observed age")
```

Equal variances

```{r}
#Homogeneity of variance
car::leveneTest(y= Age~Heart.Disease, data=heart.cleaned)
```

```{r}
#Wilcoxon rank sum test as an alternate test
wilcox.age <- wilcox.test(formula=heart.cleaned$Age~heart.cleaned$Heart.Disease, paired = FALSE)
wilcox.age
```

```{r}
#Effect size for wilcoxon rank sum test
rcompanion::wilcoxonR(x=heart.cleaned$Age, g=heart.cleaned$Heart.Disease)
```

Source Visual Outline

```{r}
#Descriptive statistics
table(heart.disease=heart.cleaned$Heart.Disease,fasting.blood.sugar.over.120=heart.cleaned$FBS.over.120)
```

H0: presence of heart disease is the same among the groups with blood sugar levels over and less than 120

HA: presence of heart disease is not the same among the groups with blood sugar levels over and less than 120

```{r}
#Chi-squared test of independence for the variables Heart disease and fasting blood sugar over 120
chisq.test(x=heart.cleaned$Heart.Disease, y=heart.cleaned$FBS.over.120)
```

```{r}
#Chi-squared examining the presence of heart disease and Fasting Blood Sugar using standardized residuals
CrossTable(heart.cleaned$Heart.Disease, heart.cleaned$FBS.over.120, expected = TRUE, prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE, chisq = TRUE, sresid = TRUE)
```

```{r}
 #Calculating the effect size using cramer's V
 library("lsr")
 cramersV(x= heart.cleaned$Heart.Disease,
          y= heart.cleaned$FBS.over.120)
```

4) Independent sample t-test for the heart disease and Blood pressure variables

Understanding the relationship between one categorical variable and one continuous variable using histogram

```{r}
#descriptive stats using histogram
heart.bp.plot <- heart.cleaned %>%
  ggplot(aes(x = BP,fill=Heart.Disease)) +
  geom_histogram(color = "white") +
  theme_minimal() +
  labs(x = "Blood pressure", y = "Patients")
heart.bp.plot
```

```{r}
#mean of bp for both presence of heart disease and no heart disease
mean_bp_heart_disease <- mean(heart.cleaned$BP[heart.cleaned$Heart.Disease == "Presence"], na.rm = TRUE)
mean_bp_no_heart_disease <- mean(heart.cleaned$BP[heart.cleaned$Heart.Disease == "Absence"], na.rm = TRUE)
print(mean_bp_heart_disease)
print(mean_bp_no_heart_disease)
```

HO: There is no difference in the mean blood pressure level between presence and absence of heart disease in the United states.

HA: There is a difference in the mean blood pressure level between presence and absence of heart disease in the United states

```{r}
#Comparing the Blood Pressure levels for the presence and absence of heart disease using t-test
t.test(formula=heart.cleaned$BP~ heart.cleaned$Heart.Disease)
```

```{r}
#Effect size using cohens D
lsr::cohensD(x=BP~Heart.Disease, data = heart.cleaned, method = "unequal")
```

Normal distribution

```{r}
#normality distribution using histogram
heart.cleaned %>%
  ggplot(aes(x=BP))+
  geom_histogram(fill ="#74639A", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="BP",
       y="Frequency")
```

Source Visual Outline

```
heart.cleaned %>%
  drop_na(BP) %>%
  ggplot(aes(sample = BP))+
  stat_qq(aes(color = "Frequency of BP", alpha = .6)) +
  facet_grid(cols = vars(Heart.Disease)) +
  geom_abline(aes(intercept = mean(x= BP),
                  slope = sd(x=BP), linetype= "Normally distributed"),
              color="gray", size=1)+
  theme_minimal()+
  labs(x="theoretical normal distribution", y="Observed BP")
```

Equal variances for each group

```{r}
#homogeneity of variance
car::leveneTest(y= BP~Heart.Disease, data=heart.cleaned)
```

Alternative test to independent sample t-test

HO: There is no difference in the ranked blood pressure values for individuals with and without heart disease in the United states. **HA: There is a difference in the ranked blood pressure values for individuals with and without heart disease in the United states.**

```{r}
#Wilcoxon test as an alternate test
wilcox.bp <- wilcox.test(formula=heart.cleaned$BP~heart.cleaned$Heart.Disease, paired = FALSE)
wilcox.bp
```

```{r}
#Effect size for wilcoxon ranksum test
rcompanion::wilcoxonR(x=heart.cleaned$BP, g=heart.cleaned$Heart.Disease)
```

5) Independent sample t-test for the heart disease and Cholesterol level variables

Understanding the relationship between one categorical variable and one continuous variable using histogram

```{r}
#descriptive stats using histogram
heart.chol.plot <- heart.cleaned %>%
  ggplot(aes(x = Cholesterol,fill=Heart.Disease)) +
  geom_histogram(color = "white") +
  theme_minimal() +
  labs(x = "Cholesterol", y = "Patients")
heart.chol.plot
```

```r
#mean of age for both presence of heart disease and no heart disease
mean_cholesterol_heart_disease <- mean(heart.cleaned$Cholesterol[heart.cleaned$Heart.Disease == "Presence"], na.rm = TRUE)
mean_cholesterol_no_heart_disease <- mean(heart.cleaned$Cholesterol[heart.cleaned$Heart.Disease == "Absence"], na.rm = TRUE)
print(mean_cholesterol_heart_disease)
print(mean_cholesterol_no_heart_disease)
```

H0: There is no difference in the mean cholesterol level between the people with and without the heart disease in the United states.

HA: There is a difference in the mean cholesterol level between the people with and without the heart disease in the United states.

```r
#Comparing the cholesterol levels for the presence and absence of heart disease using t-test
t.test(formula=heart.cleaned$Cholesterol~ heart.cleaned$Heart.Disease)
```

```r
#Effect size calculation using cohens D
lsr::cohensD(x=Cholesterol~Heart.Disease, data = heart.cleaned, method = "unequal")
```

Normal distribution

```r
#Normality histogram using histogram
heart.cleaned %>%
  ggplot(aes(x=Cholesterol))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Cholesterol",
       y="Frequency")
```

```r
heart.cleaned %>%
  drop_na(Cholesterol) %>%
  ggplot(aes(sample = BP))+
  stat_qq(aes(color = "Frequency of Cholesterol", alpha = .6)) +
  facet_grid(cols = vars(Heart.Disease)) +
  geom_abline(aes(intercept = mean(x= Cholesterol),
                  slope = sd(x=Cholesterol), linetype= "Normally distributed"),
              color="gray", size=1)+
  theme_minimal()+
  labs(x="theoretical normal distribution", y="Observed Cholesterol")
```

Assumption: Equal variance

```r
#homogeneity of variance
car::leveneTest(y= Cholesterol~Heart.Disease, data=heart.cleaned)
```

Alternative test to independent sample t-test: Wilcoxon rank-sum test

H0: There is no difference in the ranked cholesterol values for individuals with and without heart disease in the United states. **HA: There is a difference in the ranked cholesterol values for individuals with and without heart disease in the United states.**

```r
#Wilcoxon rank sum test as an alternate test
wilcox.cholesterol<- wilcox.test(formula=heart.cleaned$Cholesterol~heart.cleaned$Heart.Disease, paired = FALSE)
wilcox.cholesterol
```

p-value \< 0.05.so, reject the null hypothesis and conclude that there is a difference in the ranked cholesterol values for individuals with and without heart disease in the United states.

```r
# effect size calculating for the alternate test
rcompanion::wilcoxonR(x=heart.cleaned$Cholesterol, g=heart.cleaned$Heart.Disease)
```

6) Logistic regression model for the variables

```r
#descriptive statistics using table
table.heart <- CreateTableOne(data = heart.cleaned, strata = "Heart.Disease", vars = c("Age", "Sex", "BP", "Cholesterol", "FBS.over.120"))
```

```r
print(table.heart, nonnormal = 'age', showAllLevels = TRUE)
```

```r
#checking the levels of the heart disease
levels(as.factor(x = heart.cleaned$Heart.Disease))
```

```r
#re-leveling the levels to get the absence as a reference
heart.cleaned <- heart.cleaned %>%
  mutate(Heart.Disease = relevel(x = as.factor(Heart.Disease), ref = "Absence"))
```

```r
#rechecking the levels
levels(as.factor(x = heart.cleaned$Heart.Disease))
```

```r
#Checking the normality with linear transformations for cholesterol variable

#cube root
cube.root.cholesterol <- heart.cleaned%>%
  ggplot(aes(x=(Cholesterol)^(1/3)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Cube root of cholesterol",
       y="Frequency")

#square root
square.root.cholesterol <- heart.cleaned%>%
  ggplot(aes(x=(Cholesterol)^(1/2)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Square root of cholesterol",
       y="Frequency")

#inverse
inverse.cholesterol <- heart.cleaned%>%
  ggplot(aes(x=1/(Cholesterol)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Inverse of cholesterol",
       y="Frequency")

#log
log.cholesterol <- heart.cleaned%>%
  ggplot(aes(x=log(Cholesterol)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Log of cholesterol",
       y="Frequency")
gridExtra::grid.arrange(cube.root.cholesterol,square.root.cholesterol,inverse.cholesterol,log.cholesterol)
```

```r
#Checking the normality with linear transformations for blood pressure variable

#cube root
cube.root.bp <- heart.cleaned%>%
  ggplot(aes(x=(BP)^(1/3)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Cube root of Blood pressure",
       y="Frequency")

#square root
square.root.bp <- heart.cleaned%>%
  ggplot(aes(x=(BP)^(1/2)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Square root of Blood pressure",
       y="Frequency")

#inverse
inverse.bp <- heart.cleaned%>%
  ggplot(aes(x=1/(BP)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Inverse of Blood pressure",
       y="Frequency")

#log
log.bp <- heart.cleaned%>%
  ggplot(aes(x=log(BP)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Log of Blood pressure",
       y="Frequency")
gridExtra::grid.arrange(cube.root.bp,square.root.bp,inverse.bp,log.bp)
```

```{r}
#checking the normality with linear transformations for blood pressure variable
#cube root
cube.root.age <- heart.cleaned%>%
  ggplot(aes(x=(Age)^(1/3)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Cube root of age",
       y="Frequency")
#square root
square.root.age <- heart.cleaned%>%
  ggplot(aes(x=(Age)^(1/2)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Square root of age",
       y="Frequency")
#inverse
inverse.age <- heart.cleaned%>%
  ggplot(aes(x=1/(Age)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Inverse of age",
       y="Frequency")
#log
log.age <- heart.cleaned%>%
  ggplot(aes(x=log(Age)))+
  geom_histogram(fill ="#7463AC", col = "white")+
  facet_grid(cols=vars(Heart.Disease))+
  theme_minimal()+
  labs(x="Log of age",
       y="Frequency")
gridExtra::grid.arrange(cube.root.age,square.root.age,inverse.age,log.age)
```

Linearity:

```{r}
# make a variable of the log-odds of the outcome
logit.use.int <- log(heart.model.large$fitted.values/(1-heart.model.large$fitted.values))
# make a small data frame with the log-odds variable and the age predictor
linearity.data.int <- data.frame(logit.use.int, age.int = heart.model.large$model$Age)
# create a plot (Figure 10.14)
linearity.data.int %>%
ggplot(aes(x = age.int, y = logit.use.int))+
geom_point(aes(size = "Observation"), color = "gray", alpha = .6) +
geom_smooth(se = FALSE, aes(color = "Loess curve")) +
geom_smooth(method = lm, se = FALSE, aes(color = "linear model")) +
scale_color_manual(name = "Type of fit line",
                   values = c("dodgerblue2", "deeppink")) +
scale_size_manual(values = 1.5, name = "") +
theme_minimal() +
labs(x = "Age in years",
     y = "Log-odds of Heart diseae probability")
```

```{r}
#Make a variable of the log-odds of the outcome
logit.use.int <- log(heart.model.large$fitted.values/(1-heart.model.large$fitted.values))

#Make a small data frame with the log-odds variable and the age predictor
linearity.data.int <- data.frame(logit.use.int, bp.int = heart.model.large$model$BP)

linearity.data.int %>%
ggplot(aes(x = bp.int, y = logit.use.int))+
geom_point(aes(size = "Observation"), color = "gray", alpha = .6) +
geom_smooth(se = FALSE, aes(color = "Loess curve")) +
geom_smooth(method = lm, se = FALSE, aes(color = "linear model")) +
scale_color_manual(name = "Type of fit line",
                   values = c("dodgerblue2", "deeppink")) +
scale_size_manual(values = 1.5, name = "") +
theme_minimal() +
labs(x = "Blood pressure",
     y = "Log-odds of Heart diseae probability")
```

Multicollinearity

```{r}
#VIF
car::vif(mod = heart.model.large)
```

```{r}
#Logistic regression model using only sex variable
heart.model.small <- glm(formula = Heart.Disease ~ Sex, data = heart.cleaned, na.action = na.exclude, family=binomial("logit"))
odds.n.ends(heart.model.small)
```

##NHST for lr test

NHST Step 1: Write the null and alternate hypotheses

    HO: The larger model with the with the age, sex, and blood pressure is no better at explaining the heart disease compared to the basic model with only sex variable.
    HA: The larger model with the with the age, sex, and blood pressure is better at explaining the heart disease compared to the basic model with only sex variable.

NHST Step 2: Compute the test statistic

```{r}
# compare both the models
lmtest::lrtest(object = heart.model.small, heart.model.large)
```

NHST Step 3: Calculate the probability that your test statistic is at least as big as it is if there is no relationship (i.e., the null is true)

The probability of the test statistic was included in the output from the lrtest() function. The test statistic of $\chi 2 = 22.551$ had a p-value of <0.05.

NHST Steps 4 and 5: Interpret the probability and write a conclusion

The null hypothesis was rejected; The larger model with the with the age, sex, and blood pressure is better at explaining the heart disease compared to the basic model with only sex variable($\chi 2 = 22.551$; p = <0.05).

```{r}
# creating an example data frame to check the first model
example_data <- data.frame(Age = c(60,20),Sex= c("Female", "Male"),BP = c(183,169))
predict(object = heart.model.large, newdata = example_data, type = "response")
```

```
        1         2
0.4316819 0.2519076
```

This model considers age, sex, and blood pressure variables in predicting the presence of heart disease. According to this model, a female aged 60-years with 183mm Hg blood pressure has 18% more risk compared to a male aged 20-years with 169mm Hg blood pressure.

```{r}
#creating an example data frame to check the second model
example_data.model <- data.frame(Sex= c("Female", "Male"))
predict(object = heart.model.small, newdata = example_data.model, type = "response")
```

Conclusion:

- A logistic regression model with age, sex,and blood pressure was statistically significantly better than a baseline model at explaining heart disease [$\chi 2(3) = 47.59$; p < .001].A likelihood ratio test comparing this model to a model that included only sex variable showed that the larger model was statistically significantly better than the smaller model [$\chi 2(1) = 25.039$; p = <0.001], so the larger model was retained. This implies that the inclusion of Age and BP as additional predictors improves the model's ability to explain the variability in heart disease outcomes beyond what can be accounted for by Sex alone. In the larger model, females have approximately 92% lower odds of having heart disease compared to males (OR = 0.184, 95% CI: 0.095-0.339).The odds of heart disease were 1.06 times higher for every 1 year increase in the age(OR = 1.058, 95% CI: 1.026 - 1.093). For every unit increase in blood pressure, the odds of heart disease increase by approximately 1.02% (OR = 1.0165; 95% CI: 1.001-1.032). Assumption checking revealed a possible problem with the linearity of the age and blood pressure predictor. The other assumptions were met. The model with Age, Sex, and Blood Pressure as predictors had a slightly higher overall accuracy of 68.9% but lower sensitivity (62.5%) compared to the Sex-only model.