Heart Disease Analysis

Ali Agharezakashi
1008264049
University of Toronto
STA312
Professor: Nishan Mudalige

Table of Content

Introduction

Overview

Heart disease is the second leading cause of death in Canada. Additionally, worldwide heart disease is one of the biggest and most relevant topics. It is a topic of interest in many fields and people of different backgrounds including many doctors and researchers. Unfortunately, there is a lot of vagueness regarding the factors that directly affect or are related to heart diseases. This report will dive deeper into the topic to seek some clarification regarding the topic and answer the following 4 goal questions.

Goals

1. Is age a significant contributing factor to cholesterol levels?
2. Is resting blood pressure needed to accurately predict and draw conclusions about a patient's cholesterol levels?
3. Is a patient's cholesterol levels a significant enough indicator of the possibility of heart disease?
4. Are there any indicators significantly contributing to a patient's low blood pressure over healthy?

Description

Heart disease remains a significant health challenge in Canada and globally, driven by lifestyle changes and modernization. Contributing factors include insufficient exercise, unhealthy diets, and increasing stress levels. Efforts to understand and address heart disease focus on prevention, early detection, and treatment by identifying key predictive factors and indicators. This report aims to explore these hypothesized factors to better understand their role in heart disease among Canadians and worldwide populations.

This topic is well-documented, providing a solid foundation for further exploration based on established findings. According to the Canadian Chronic Disease Surveillance System, heart disease affects approximately 2.6 million Canadian adults aged 20 and older, with men at higher risk and often diagnosed earlier than women. Adults with heart disease face nearly three times the mortality risk compared to those without, with even higher risks for individuals who have experienced heart attacks or heart failure. However, from 2000 to 2018, new diagnoses and mortality rates have declined, reflecting advancements in prevention and treatment. Modifiable risk factors, including smoking, physical inactivity, poor diet, and excessive alcohol consumption, along with early management of conditions such as high blood pressure, and diabetes, can significantly reduce the risk of heart disease (Public Health Agency of Canada, 2022).

Description of the Datasets

The data sets include records and indicators related to heart disease and the diagnosis of heart disease. The datasets are derived from

- Kaggle
- UC Irvine Machine Learning Repository
- Statistics Canada

The data sets include variables such as age, sex, cholesterol levels, diagnosis of heart disease, etc.

Kaggle Data Set

This dataset originates from the UC Irvine Machine Learning Repository and combines data from four major sources: Cleveland, Hungary, Switzerland, and Long Beach. While the full database includes 76 attributes, research studies typically focus on a recommended subset of 14 attributes. The version used here is a cleaned Kaggle dataset derived from the original UC Irvine database, containing these 14 attributes. This multivariate dataset is not only valuable for understanding heart disease but also for predicting its presence and severity using the 14 indicators described in the table below.

Variable Name	Type	Description	Units(if applicable)
age	Integer		years
origine		Place of Study	
sex	Categorical		male/female
id		Unique Id Number	
cp	Categorical	Chest Pain Type	typical angina, atypical angina, non-anginal, asymptomatic
trestps	Integer	Resting Blood Pressure	mm Hg
chol	Integer	Serum Cholesterol	mg/dl
fbs	Categorical	(if fasting blood sugar > 120 mg/dl)	
restecg	Categorical	Resting Electrocardiographic Results	normal, stt abnormality, lv hypertrophy
thalach	Integer	Maximum Heart Rate Achieved	
exang	Categorical	Exercise-Induced Angina	
oldpeak	Integer	ST depression induced by exercise relative to rest	
slope	Categorical	The Slope of the Peak Exercise ST Segment	
ca	Integer	number of major vessels	0,1,2,3
thal	Categorical	Thalassemia	Normal, fixed defect, reversible defect
num	Integer	Diagnosis of Heart Disease	0: No diagnosis 1,2,3,4: Stages of Heart Disease

UC Irvine Data Set

This database includes multiple versions of the dataset collected from the four locations mentioned in the Kaggle dataset. However, due to the availability of both processed and unprocessed versions and reported issues with the Cleveland dataset, it was excluded from computational analyses in this report. Nevertheless, the dataset served as a valuable reference, providing detailed insights into the variables, their origins, and the data collection methods

The authors of the databases have requested that any publications resulting from the use of the data include the names of the principal investigator responsible for the data collection at each institution. They would be:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Government of Canada Data Set

The dataset obtained from the Government of Canada was not used for computational analysis. However, it provided valuable information and insights relevant to this report. Its data visualizations and statistics helped shape the direction of the study and supported various explanations throughout the report.

Methodology

To address the four proposed questions, a suitable dataset was selected from the collections provided by the hospitals and universities listed under the UC Irvine dataset section. These datasets were merged by the UC Irvine Machine Learning Repository and later formatted and posted on Kaggle by Redwan Sony (UCI Heart Disease Data, 2020). This dataset was used for analysis to explore the research questions. For each question, a null hypothesis was formulated to guide the data analysis and provide structured answers.

Hypothesis and Test Statistic

- Hypothesis 1: Age is significant in predicting cholesterol levels.
    - Regression Type: Multiple Linear Regression
    - Dependent Variable: Cholesterol level (continuous)
    - Null Hypothesis (H_0): Age does not significantly predict cholesterol levels($\beta_{age} = 0$)
    - Alternative Hypothesis (H_1): Age significantly predicts cholesterol levels. ($\beta_{age} \neq 0$)
    - Test Statistic: t-statistic for the age predictor in the linear regression model.
    - Interpretation: If the t-test is significant, H_0could be rejected, concluding that age significantly affects cholesterol levels.
- Hypothesis 2: Resting blood pressure is not needed to accurately predict cholesterol levels
    - Regression Type: Multiple Linear Regression (with nested model comparison)
    - Dependent Variable: Cholesterol level (continuous)
    - Models
        - Nested model without blood pressure as a predictor.
        - Full model with blood pressure as a predictor.

- Null Hypothesis (H_0): The model without blood pressure predicts cholesterol as accurately as the full model (the addition of blood pressure does not improve model fit).
- Alternative Hypothesis (H_1): Including blood pressure improves the accuracy of the model for predicting cholesterol.
- Test Statistic: F-statistic.
- Interpretation: If the F-test is significant, H_0could be rejected concluded that blood pressure improves the model's predictive accuracy for cholesterol.

- Hypothesis 3: Cholesterol level is significant in predicting heart disease
  - Regression Type: Binary Logistic Regression
  - Dependent Variable: Diagnosis of heart disease (binary: Yes or No)(Data will be transformed from 0,1,2,3,4 to 0: No, (1,2,3,4): Yes)
  - Null Hypothesis (H_0): Cholesterol levels do not significantly predict heart disease ($\beta_{chol} = 0$).
  - Alternative Hypothesis (H_1): Cholesterol levels significantly predict heart disease ($\beta_{chol} \neq 0$).
  - Test Statistics: chi-square test for the cholesterol predictor.
  - Interpretation: If the chi-square test is significant,H_0could be rejected, concluding that cholesterol levels significantly impact the likelihood of a heart disease diagnosis.
- Hypothesis 4: None of the predictors affect a patient having low blood pressure over healthy.
  - Regression Type: Ordinal Regression
  - Dependent Variable: Cholesterol level (categorical: Low, Healthy, High)
  - Reference Category: Healthy
  - Predictors: Any predictors in the model, e.g., age, cholesterol level, etc.
  - Null Hypothesis (H_0): None of the predictors significantly differentiate between low and healthy cholesterol levels ($\forall\, i, \beta_i^{low} = 0$).
  - Alternative Hypothesis (H_1): At least one predictor significantly differentiates between low and healthy cholesterol levels($\exists\, i,\ \beta_i^{low} \neq 0$).
  - Test Statistic: chi-square test for each predictor in the multinomial logistic regression.
  - Interpretation: If the chi-square test for any predictor is significant, reject H_0 ,indicating that at least one predictor significantly differentiates low cholesterol from healthy levels.

Analysis & Results

Cleaning the data

To conduct the analysis, the dataset was modified and processed. The original dataset included a column indicating the location of data collection (e.g., Cleveland), which was deemed irrelevant to the analysis and was therefore removed (Appendix Figure 1).

Additionally, the Kaggle dataset contained unique patient ID numbers, which were unnecessary for the study and could interfere with model development, so this column was also removed.

The resulting processed dataset, named *uci_processed*, was saved to the working directory for use in all subsequent analyses. Furthermore, Variance Inflation Factor (VIF) analysis confirmed the absence of multicollinearity among the predictors (Figure 2).

```
vif(multilinear_model_basic)

##              GVIF Df GVIF^(1/(2*Df))
## id       1.379707  1        1.174609
## age      1.484452  1        1.218381
## sex      1.324401  1        1.150826
## dataset  1.313835  2        1.070620
## cp       1.837045  3        1.106675
## trestbps 1.239843  1        1.113482
## fbs      1.129259  1        1.062666
## restecg  1.243953  2        1.056090
## thalch   1.803497  1        1.342943
## exang    1.460274  1        1.208418
## oldpeak  1.993912  1        1.412060
## slope    2.008075  2        1.190406
## ca       1.625322  1        1.274881
## thal     1.809428  2        1.159806
## num      2.424352  1        1.557033
```

Figure 2: VIF Table

Hypothesis 1

In order to prove or disprove the first hypothesis a linear model was fitted into the data set to check if age is a significant factor contributing toward cholesterol levels.(Appendix Figure 3).

The following summary was achieved(Appendix Figure 4).

Result:

Based on the given P value from the t-test given in Figure 4, we have

$Pr(>|t|) = 0.01688 < 0.05 \Rightarrow$ It was concluded that age is a significant factor contributing to the value of a patient's cholesterol levels. $\Rightarrow$ H_0 could be rejected and H_1 could be concluded($\beta_{age} \neq 0$).

In order to visually represent the validity the model a relationship plot(Figure 7) and a Q-Q plot were made(Figure6)

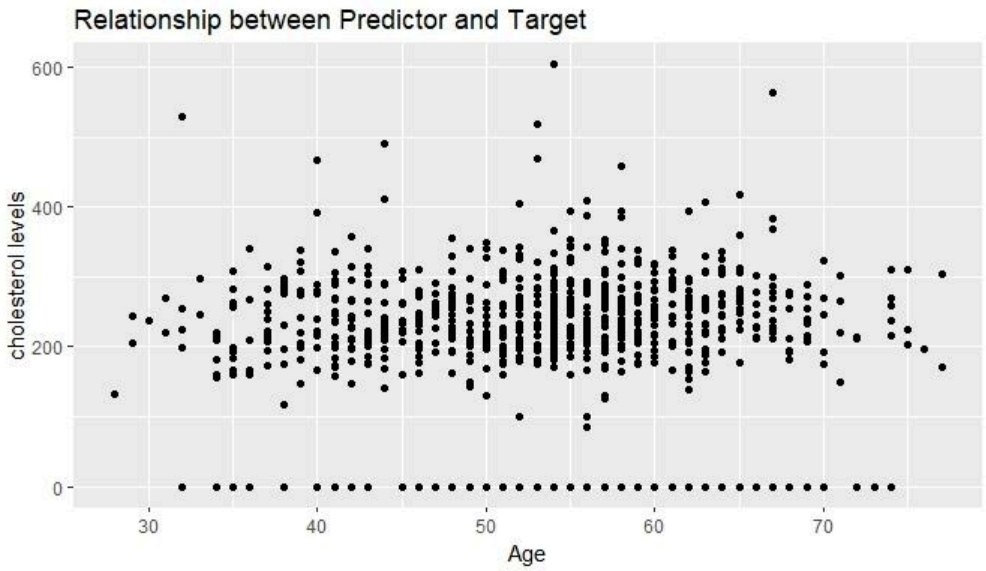

Figure 5: Relationship between Age and Cholesterol levels

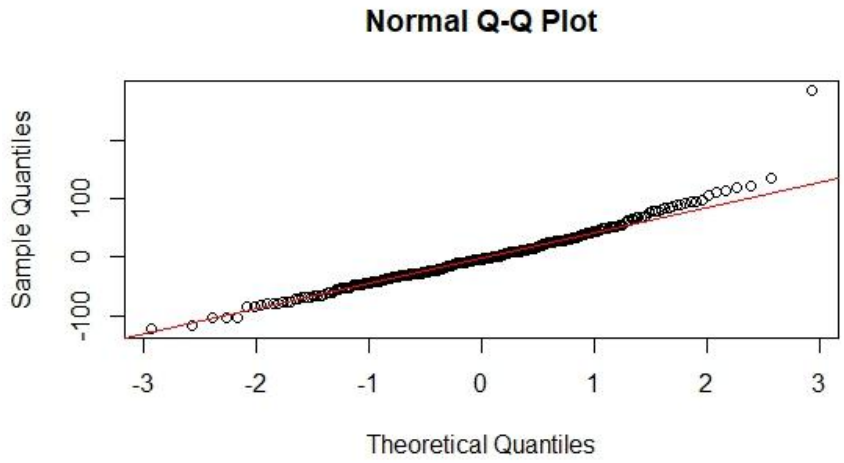

Figure 6: Normal Q-Q Plot

Hypothesis 2

In order to prove or disprove the second hypothesis a full model and a nested model without the resting blood pressure was created(Appendix Figure7 & Figure8).

After the models were created a model comparison was made using Anova and *F*-test to see if a reduced model without the resting blood pressure could accurately predict cholesterol levels as the full model(Figure9)

```
model_comparison <- anova(nested_model, full_model)
print(model_comparison)

## Analysis of Variance Table
##
## Model 1: uci_processed$chol ~ (id + age + sex + cp + trestbps + fbs +
##     restecg + thalch + exang + oldpeak + slope + ca + thal +
##     num) - trestbps
## Model 2: uci_processed$chol ~ id + age + sex + cp + trestbps + fbs +
restecg +
##     thalch + exang + oldpeak + slope + ca + thal + num
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    280 703741
## 2    279 701050  1    2691.2 1.071 0.3016
```

Figure 9 : Full vs Nested Model Comparison

Result:

According to Figure 9 a full model is not required and we can accurately predict cholesterol levels without knowing resting blood pressure as we have

$Pr(> F) = 0.3016 > 0.05 \Rightarrow$ we cannot reject H_0.

Hypothesis 3

As mentioned earlier. The data set used for computation and analysis has a variable *num* which could have values(0,1,2,3,4). 0 means the lack of presence of heart disease. (1,2,3,4) means the presence of heart disease in the patient and the numbers indicate different stages of heart disease. The goal questions and hypothesized answer is dealing with a binary answer of yes and no. Hence a transformed list was created to store the values of num as a binary 0,1(0:No,1:Yes)(Appendix Figure 10).

Subsequently a logistic binary regression was written with $Binary\_heartdisease$ as the dependent variable and variable *num* was excluded from the predictors(Appendix Figure 11).

Result:

An analysis of the model showed that contrary to popular belief cholesterol level is not a significant enough of a factor in predicting heart disease(Appendix Figure 12).

Based on Figure 12 $Pr(> |Z|)_{chol} = 0.287409 > 0.05 \Rightarrow$ fail to reject $H_0(\beta_{chol} = 0)$

Hypothesis 4

Firstly the value for target value ($chol$) which represents the cholesterol levels of the patients is continuous and numerical. However in order to be able to prove the final hypothesis to be true or untrue they should be transformed to ranked categorical values. Hence a new list of cholesterols was created named ($ordinal\_chol$)(Appendix Figure 13)

After that Healthy was set as the reference state for $chol$(Appendix Figure 13) and model summary was obtained(Appendix Figure 14).For better analysis and visualization purposes the coefficients were isolate along with their *P* values(Appendix Figure 15).

Results of the P values were created and significant ones were isolated(Figure 16).

```
print(result)

##                          Coefficients   P_Values
## age                       0.013227946 0.37468295
## sexMale                  -0.544786589 0.05653628
## cpatypical angina        -0.015393143 0.96793647
## cpnon-anginal            -0.004532843 0.98838157
## cptypical angina         -0.497083078 0.28647250
## trestbps                  0.008714322 0.23384557
## fbsTRUE                   0.042691748 0.90241839
## restecgnormal            -0.582079580 0.01546531
## restecgst-t abnormality -0.575716338 0.57549395
## thalch                   -0.000811560 0.89892515
## exangTRUE                 0.363822966 0.21780339
## oldpeak                  -0.189481455 0.15622211
## slopeflat                -0.375974765 0.45352883
## slopeupsloping           -0.464680667 0.40158232
## ca                        0.154929475 0.28098271
## thalnormal                0.635953373 0.21899545
## thalreversable defect     0.636318427 0.19200514
## Healthy|Low               0.427664380 0.80441794
## Low|High                  1.170858014 0.49805753
```

```
print(significant_results)

##                Coefficients   P_Values
## restecgnormal   -0.5820796 0.01546531
```

Figure 16: Significant Predictors

Result:

Hence based on the summary table given in Figure 16 and the $significantresults$ it was concluded that the H_0could be rejected meaning there exist a predictor that affect a patient having low blood pressure over healthy and that predictor $restecgnormal$ has P_values $0.01546531 < 0.05 \Rightarrow \beta_{restecgnormal} \neq 0$. There exists a coefficient that is not 0.

Conclusion

In conclusion, hypotheses one and four were rejected, while hypotheses two and three were not.

Following modifications to prepare the dataset for computation, it was confirmed that the predictors were uncorrelated, and the Q-Q plot indicated an approximately normal distribution for hypothesis one. Analysis revealed that age significantly influences cholesterol levels, aligning with the conventional belief that older individuals are more likely to have elevated cholesterol, increasing health risks.

For hypothesis two, the null hypothesis could not be rejected, indicating that resting blood pressure does not significantly enhance the predictive accuracy of cholesterol levels when other predictors are included. However, this does not rule out a potential direct or indirect relationship between resting blood pressure, cholesterol levels, and heart disease, warranting further targeted investigation.

The analysis for hypothesis three showed that cholesterol levels do not significantly impact the likelihood of developing heart disease. Notably, being male was a more significant factor, consistent with Canadian

health data stating that men are twice as likely as women to suffer a heart attack (Public Health Agency of Canada, 2022). It is possible that analyzing the stages of heart disease might yield different insights.

Finally, the results for the fourth hypothesis confirmed that at least one predictor was significant, with $restecgnormal$ which represents the normal mode of resting electrocardiographic results.

Appendix

```
> uci_processed <- select(heart_disease_uci, -dataset)
> View(uci_processed)
> save(uci_processed, file = "C:/Users/alini/Desktop/University/2024/Fall/STA312/
Project/uci_processed.RData")
```

Figure 1: Location Column Removal

```
multilinear_model_basic <- lm(heart_disease_uci$chol ~ . ,data = heart_disease_uci)
```

Figure 3 : Multilinear Model

```
summary(multilinear_model_basic)

##
## Call:
## lm(formula = heart_disease_uci$chol ~ ., data = heart_disease_uci)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.587  -30.334   -3.078   27.685  283.776
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              137.71810   46.34734   2.971  0.00322 **
## id                        -0.04822    0.03544  -1.361  0.17468
## age                        0.94011    0.39035   2.408  0.01668 *
## sexMale                  -22.65493    7.11932  -3.182  0.00163 **
## datasetHungary            16.66600   52.65605   0.317  0.75186
## datasetVA Long Beach    -106.49853   54.80140  -1.943  0.05299 .
## cpatypical angina          2.41452    9.64144   0.250  0.80244
## cpnon-anginal             -2.72623    8.14164  -0.335  0.73799
## cptypical angina         -10.52070   12.34668  -0.852  0.39489
## trestbps                   0.16616    0.18151   0.915  0.36077
## fbsTRUE                   -2.09408    8.74692  -0.239  0.81097
## restecgnormal            -13.71481    6.18117  -2.219  0.02731 *
## restecgst-t abnormality    3.52670   26.53858   0.133  0.89438
## thalch                     0.23917    0.16804   1.423  0.15578
## exangTRUE                  7.52688    7.41647   1.015  0.31105
## oldpeak                    0.30596    3.51341   0.087  0.93067
## slopeflat                 12.24410   12.79485   0.957  0.33942
## slopeupsloping            12.85355   14.12531   0.910  0.36363
## ca                         3.20769    3.93456   0.815  0.41562
## thalnormal                12.88730   13.65052   0.944  0.34595
## thalreversable defect     19.51351   13.08207   1.492  0.13694
## num                        0.23984    3.66115   0.066  0.94782
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Figure 4 : Multilinear Model Summary

```
summary(full_model)

##
## Call:
## lm(formula = uci_processed$chol ~ ., data = uci_processed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.375  -31.235   -2.686   26.971  286.026
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              143.47524   46.42899   3.090  0.00220 **
## id                        -0.07103    0.03225  -2.202  0.02847 *
## age                        0.89412    0.38990   2.293  0.02258 *
## sexMale                  -23.22211    7.12995  -3.257  0.00127 **
## cpatypical angina          2.84544    9.67213   0.294  0.76883
## cpnon-anginal             -2.66308    8.17102  -0.326  0.74473
## cptypical angina          -9.87490   12.37976  -0.798  0.42574
## trestbps                   0.18749    0.18116   1.035  0.30161
## fbsTRUE                   -2.32379    8.77836  -0.265  0.79142
## restecgnormal            -13.55198    6.20279  -2.185  0.02973 *
## restecgst-t abnormality    6.26724   26.57235   0.236  0.81372
## thalch                     0.22521    0.16757   1.344  0.18004
## exangTRUE                  6.90147    7.43306   0.928  0.35396
## oldpeak                   -0.11081    3.51747  -0.032  0.97489
## slopeflat                 12.28723   12.84121   0.957  0.33947
## slopeupsloping            13.06301   14.17681   0.921  0.35762
## ca                         3.66221    3.93938   0.930  0.35336
## thalnormal                12.45754   13.68928   0.910  0.36360
## thalreversable defect     18.41864   13.08677   1.407  0.16041
## num                        0.54225    3.65893   0.148  0.88229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.13 on 279 degrees of freedom
##   (621 observations deleted due to missingness)
## Multiple R-squared:  0.1475, Adjusted R-squared:  0.08948
## F-statistic: 2.541 on 19 and 279 DF,  p-value: 0.0005177
```

Figure 7 : Full Multilinear Model

```
summary(nested_model)

##
## Call:
## lm(formula = uci_processed$chol ~ . - trestbps, data = uci_processed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.424  -33.411   -3.872   27.453  280.664
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              161.38168   43.08993   3.745 0.000219 ***
## id                        -0.07016    0.03225  -2.176 0.030408 *
## age                        0.99153    0.37842   2.620 0.009268 **
## sexMale                  -24.18474    7.06992  -3.421 0.000717 ***
## cpatypical angina          3.09186    9.67042   0.320 0.749417
## cpnon-anginal             -2.57398    8.17160  -0.315 0.753004
## cptypical angina          -8.19941   12.27500  -0.668 0.504699
## fbsTRUE                   -1.08736    8.69778  -0.125 0.900601
## restecgnormal            -14.13457    6.17798  -2.288 0.022889 *
## restecgst-t abnormality    6.32611   26.57566   0.238 0.812023
## thalch                     0.24223    0.16678   1.452 0.147517
## exangTRUE                  7.15196    7.43006   0.963 0.336594
## oldpeak                    0.22978    3.50249   0.066 0.947739
## slopeflat                 11.76347   12.83286   0.917 0.360105
## slopeupsloping            13.03289   14.17857   0.919 0.358784
## ca                         3.31254    3.92536   0.844 0.399456
## thalnormal                11.18890   13.63601   0.821 0.412606
## thalreversable defect     17.94252   13.08034   1.372 0.171250
## num                        0.86476    3.64610   0.237 0.812695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.13 on 280 degrees of freedom
##   (621 observations deleted due to missingness)
## Multiple R-squared:  0.1443, Adjusted R-squared:  0.08925
## F-statistic: 2.622 on 18 and 280 DF,  p-value: 0.0004362
```

Figure 8 : Nested Model

```
print(Binary_heartdisease)
```

```
##   [1] 0 1 1 0 0 0 1 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 0 0 0 1 0 1 1 0
0 0 1
##  [38] 1 1 0 1 0 0 0 1 1 0 1 0 0 0 0 1 0 1 1 1 1 0 0 1 0 1 0 1 1 1 0 1 1 0
1 1 1
##  [75] 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 1 0 1 1
1 1 1
## [112] 1 0 1 1 0 0 0 1 1 1 1 0 1 1 0 1 1 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 1 0
1 1 0
## [149] 0 0 0 0 0 1 1 1 1 1 1 0 0 1 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 0 1 1
0 0 1
## [186] 0 0 1 1 1 0 1 1 1 0 1 0 0 0 1 0 0 0 0 0 1 1 1 0 1 0 1 0 1 1 0 0 0 0
0 0 0
## [223] 0 1 1 0 0 0 1 1 0 1 1 0 0 1 1 1 0 0 0 0 0 1 0 1 1 1 1 0 0 1 0 0 0 0
0 0 0
## [260] 1 0 1 0 0 1 1 1 1 1 0 1 0 1 0 1 0 0 0 1 0 1 0 1 0 1 1 1 0 0 0 1 0 1
1 1 0
## [297] 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0
## [334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0
```

```
## [371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0
## [408] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0
## [445] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0
## [482] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
## [519] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
## [556] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
## [593] 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
## [630] 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1
1 1 1
## [667] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1
## [704] 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 0 0 1 0 1 1 0 1 1 1 1 0 1
1 0 0
## [741] 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 1 0 1 0 1 0 1 0
1 1 1
## [778] 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1
0 1 1
## [815] 0 1 0 1 1 0 1 1 1 1 0 1 1 1 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1
0 0 1
## [852] 1 1 0 1 0 1 1 0 1 0 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 0
## [889] 1 1 1 0 1 1 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 0 0 1 1 1 1 1 0 1 0 1
```

Figure 10: Numerical to Binary Transformation of num

```
logistic_binary_model <- glm(Binary_heartdisease ~ . -num, data = uci_processed, family = binomial())
```

Figure 11: Logistic Binary Model

```
ordinal_chol <- cut(uci_processed$chol,breaks = c(-Inf, 199, 239, Inf),labels = c("Low", "Healthy", "High"))
```

Figure 12: ordinal_chol

```
summary(logistic_binary_model)

##
## Call:
## glm(formula = Binary_heartdisease ~ . - num, family = binomial(),
##     data = uci_processed)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.890740   2.847494  -1.015 0.310017
## age                      -0.013708   0.024730  -0.554 0.579367
## sexMale                   1.550025   0.530800   2.920 0.003498 **
## cpatypical angina        -0.845413   0.560658  -1.508 0.131582
## cpnon-anginal            -1.850092   0.501372  -3.690 0.000224 ***
## cptypical angina         -2.101023   0.666796  -3.151 0.001628 **
## trestbps                  0.024416   0.011254   2.170 0.030036 *
## chol                      0.004206   0.003954   1.064 0.287409
## fbsTRUE                  -0.595202   0.609262  -0.977 0.328607
## restecgnormal            -0.469487   0.384077  -1.222 0.221566
## restecgst-t abnormality   0.312786   2.438643   0.128 0.897941
## thalch                   -0.018049   0.011104  -1.625 0.104067
## exangTRUE                 0.718987   0.439930   1.634 0.102191
## oldpeak                   0.361006   0.230521   1.566 0.117338
## slopeflat                 0.647394   0.850593   0.761 0.446592
## slopeupsloping           -0.516696   0.922437  -0.560 0.575382
## ca                        1.309933   0.279769   4.682 2.84e-06 ***
## thalnormal                0.031320   0.789773   0.040 0.968367
## thalreversable defect     1.432948   0.774593   1.850 0.064323 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 413.03  on 298  degrees of freedom
## Residual deviance: 191.97  on 280  degrees of freedom
##   (621 observations deleted due to missingness)
## AIC: 229.97
##
## Number of Fisher Scoring iterations: 6
```

Figure 12: Logistic Binary Model Summary

```
ordinal_chol <- relevel(ordinal_chol, ref = "Healthy")
```

Figure 13: Putting Healthy as Reference

```
summary(ordinal_model)

## Call:
## polr(formula = ordinal_chol ~ . - num - chol, data = uci_processed,
##     Hess = TRUE)
##
## Coefficients:
##                              Value Std. Error  t value
## age                       0.0132279   0.014901  0.88774
## sexMale                  -0.5447866   0.285695 -1.90688
## cpatypical angina        -0.0153931   0.382947 -0.04020
## cpnon-anginal            -0.0045328   0.311278 -0.01456
## cptypical angina         -0.4970831   0.466354 -1.06589
## trestbps                  0.0087143   0.007320  1.19051
## fbsTRUE                   0.0426917   0.348200  0.12261
## restecgnormal            -0.5820796   0.240400 -2.42130
## restecgst-t abnormality  -0.5757163   1.028104 -0.55998
## thalch                   -0.0008116   0.006389 -0.12702
## exangTRUE                 0.3638230   0.295217  1.23239
## oldpeak                  -0.1894815   0.133636 -1.41789
## slopeflat                -0.3759748   0.501604 -0.74954
## slopeupsloping           -0.4646807   0.553984 -0.83880
## ca                        0.1549295   0.143704  1.07811
## thalnormal                0.6359534   0.517370  1.22920
## thalreversable defect     0.6363184   0.487724  1.30467
##
## Intercepts:
##              Value   Std. Error t value
## Healthy|Low  0.4277  1.7270     0.2476
## Low|High     1.1709  1.7281     0.6775
##
## Residual Deviance: 573.2473
## AIC: 611.2473
```

Figure 14: Ordinal Model Summary

```
coefficients <- summary(ordinal_model)$coefficients
p_values <- 2 * (1 - pnorm(abs(coefficients[, "t value"])))
```

Figure 15: P Values

References

UCI heart disease data. (2020, September 23). https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data/data

Heart disease in Canada: Highlights from the Canadian Chronic Disease Surveillance System - Open Government Portal. (n.d.). https://open.canada.ca/data/en/dataset/73808a4c-bcf4-4d6b-a9d1-301c31e32e14

Heart disease - UCI Machine Learning Repository. (n.d.). https://archive.ics.uci.edu/dataset/45/heart+disease

Public Health Agency of Canada. (2022, July 28). *Heart disease in Canada*. Canada.ca. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/heart-disease-canada.html

Heart Risk & prevention. (n.d.). Heart and Stroke Foundation of Canada. https://www.heartandstroke.ca/heart-disease/risk-and-prevention

Grolemund, G. (2014, July 16). *Introduction to R markdown*. https://rmarkdown.rstudio.com/articles_intro.html

Heart disease dataset. (2019, June 6). https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Public Health Agency of Canada. (2024, January 16). *Canadian Chronic Disease Surveillance System (CCDSS) — Canada.ca*. https://health-infobase.canada.ca/ccdss/data-tool/Index?G=00&V=9&M=5&Y=2017

Detrano, R., Jánosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). *International application of a new probability algorithm for the diagnosis of coronary artery disease.*

https://www.semanticscholar.org/paper/International-application-of-a-new-probability-for-Detrano-J%C3%A1nosi/a7d714f8f87bfc41351eb5ae1e5472f0ebbe0574

.