WHAT IS YOUR HEART RATE TELLING YOU ?

image

Heart disease and potential risk factors

Annually, countless individuals are diagnosed with various forms of heart disease, which stands as the leading cause of death for both genders across the globe, including in the United States. Through statistical studies, several factors contributing to the likelihood of developing heart disease have been pinpointed. These include but are not limited to age, blood pressure, cholesterol levels, diabetes, hypertension, genetic predisposition, obesity, and a sedentary lifestyle. In this project, I aim to apply statistical analyses and regression methodologies on the Cleveland heart disease dataset, focusing specifically on the relationship between the maximum heart rate achievable during physical activity and its correlation with an increased risk of heart disease.

The Data

Available on `Cleveland_hd.csv` | Column | Type | Description |
|------------|------|-------------------------| |`age` | continuous | age in years | |`sex` | discrete | 0=female 1=male | |`cp`| discrete | chest pain type: 1=typical angina, 2=atypical angina, 3=non-anginal pain, 4=asymptom | |`trestbps`| continuous | resting blood pressure (in mm Hg) | |`chol`| continuous | serum cholesterol in mg/dl | |`fbs`| discrete | fasting blood sugar>120 mg/dl: 1=true 0=False | |`restecg`| discrete | result of electrocardiogram while at rest are represented in 3 distinct values 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 Estes' criteria (Nominal) | |`thalach`| continuous | maximum heart rate achieved | |`exang`| discrete | exercise induced angina: 1=yes 0=no | |`oldpeak`| continuous | depression induced by exercise relative to rest | |`slope`| discrete | the slope of the peak exercise segment: 1=up sloping 2=flat, 3=down sloping |`ca`| continuous | number of major vessels colored by fluoroscopy that ranged between 0 and 3 | |`thal`| discrete | 3=normal 6=fixed defect 7=reversible defect | |`class`| discrete | diagnosis classes: 0=no presence 1=minor indicators for heart disease 2=>1 3=>2 4=major indicators for heart disease|

```
# Loading the necessary packages
install.packages("caret") #caret package for model building
library(caret)

# Reading the datasets Cleveland_hd.csv into hd_data
hd_data <- read.csv("Cleveland_hd.csv")

# taking a look at the first 5 rows of hd_data
head(hd_data, 5)
```

```
Installing caret [6.0-94] ...
      OK [linked cache]
```


```
{"table":{"data":{"age":[63,67,67,37,41],"ca":[0,3,2,0,0],"chol":
[233,286,229,250,204],"class":[0,2,1,0,0],"cp":[1,4,4,3,2],"exang":
[0,1,1,0,0],"fbs":[1,0,0,0,0],"index":["1","2","3","4","5"],"oldpeak":
[2.3,1.5,2.6,3.5,1.4],"restecg":[2,2,2,0,2],"sex":[1,1,1,1,0],"slope":
[3,2,2,3,1],"thal":[6,3,7,3,3],"thalach":
[150,108,129,187,172],"trestbps":[145,160,120,130,130]},"schema":
{"fields":[{"name":"index","type":"string"},
{"name":"age","type":"integer"},{"name":"sex","type":"integer"},
{"name":"cp","type":"integer"},{"name":"trestbps","type":"integer"},
{"name":"chol","type":"integer"},{"name":"fbs","type":"integer"},
{"name":"restecg","type":"integer"},
{"name":"thalach","type":"integer"},{"name":"exang","type":"integer"},
{"name":"oldpeak","type":"float"},{"name":"slope","type":"integer"},
{"name":"ca","type":"integer"},{"name":"thal","type":"integer"},
{"name":"class","type":"integer"}],"pandas_version":"0.20.0","primaryK
ey":["index"]}},"total_rows":5,"truncation_type":null}
```


Converting diagnosis class into outcome variable

We noticed that the outcome variable class has more than two levels. According to the codebook, any non-zero values can be coded as an "event." Let's create a new variable called hd to represent a binary 1/0 outcome. There are a few other categorical/discrete variables in the dataset. Let's also convert sex into a 'factor' for next step analysis. Otherwise, R will treat this as continuous by default.

```
# Loading the tidyverse package
library(tidyverse)

# Recoding the 'hd' variable in hd_data
hd_data <- mutate(hd_data, hd = ifelse(class > 0, 1, 0))

# Recoding 'sex' as a factor and saving the result back to hd_data
hd_data <- mutate(hd_data, sex = factor(sex, levels = 0:1, labels =
c("Female", "Male")))
```

Identifying important clinical variables

Now, let's use statistical tests to see which predictors are related to heart disease. We can explore the associations for each variable in the dataset. Depending on the type of the data (i.e., continuous or categorical), we use t-test or chi-squared test to calculate the p-values. Recall, t-test is used to determine whether there is a significant difference between the means of two

groups (e.g., is the mean age from group A different from the mean age from group B?). A chi-squared test for independence compares the equivalence of two proportions.

```r
# Does sex have an effect? Sex is a binary variable in this dataset,
# so the appropriate test is chi-squared test
hd_sex <- chisq.test(hd_data$sex, hd_data$hd)

# Since 22.043 is much greater than the critical value of 3.841,and
the the p value is significantlly less than 0.05 ,  we can conclude
that the observed association between sex and heart disease status is
statistically significant.

# Does age have an effect? Age is continuous, so we use t-test here
hd_age <- t.test(hd_data$age ~ hd_data$hd)

#The t-test results indicate that there is a statistically significant
difference in average age between individuals without heart disease
(mean age = 52.59 years) and those with heart disease (mean age =
56.63 years), with a p-value of 7.061e-05, suggesting that age may
have an effect on the likelihood of developing heart disease.

# What about thalach: maximum heart rate one can achieve during
exercise?
hd_heartrate <- t.test(hd_data$thalach ~ hd_data$hd)

#The t-test results reveal a statistically significant difference in
the maximum heart rate achieved during exercise between individuals
without heart disease (mean = 158.38 bpm) and those with heart disease
(mean = 139.26 bpm), with a p-value of 9.106e-14. This suggests that a
higher maximum heart rate during exercise is associated with a lower
likelihood of having heart disease.

# Print the results to see if p<0.05.
print(hd_sex)
print(hd_age)
print(hd_heartrate)
```

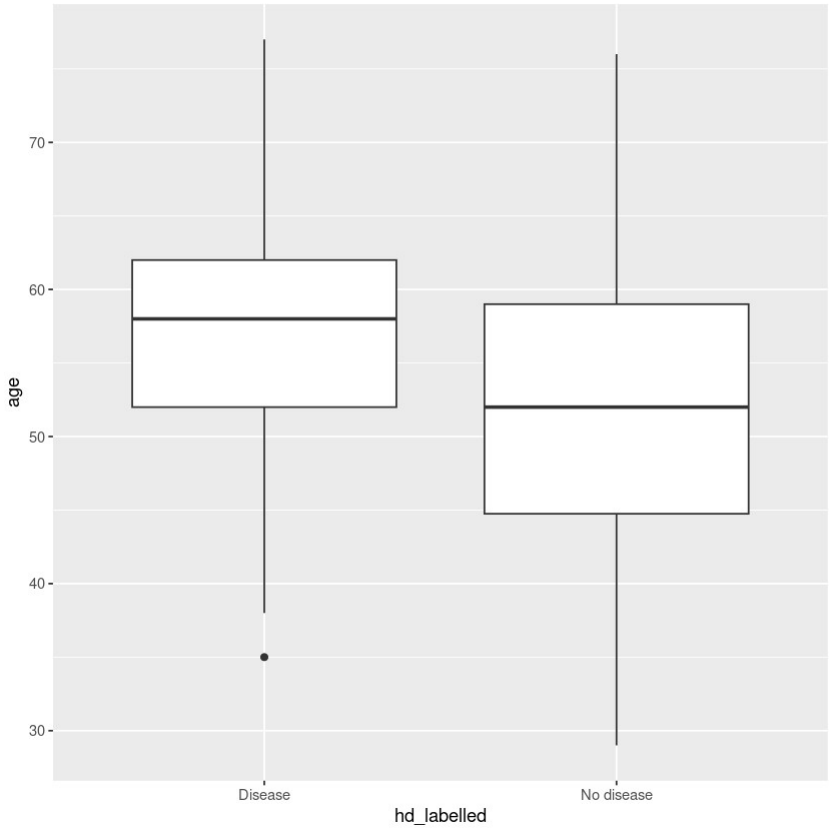
```
     Pearson's Chi-squared test with Yates' continuity correction

data:  hd_data$sex and hd_data$hd
X-squared = 22.043, df = 1, p-value = 2.667e-06


     Welch Two Sample t-test

data:  hd_data$age by hd_data$hd
t = -4.0303, df = 300.93, p-value = 7.061e-05
alternative hypothesis: true difference in means between group 0 and
group 1 is not equal to 0
95 percent confidence interval:
```

```
 -6.013385 -2.067682
sample estimates:
mean in group 0 mean in group 1
       52.58537        56.62590


     Welch Two Sample t-test

data:  hd_data$thalach by hd_data$hd
t = 7.8579, df = 272.27, p-value = 9.106e-14
alternative hypothesis: true difference in means between group 0 and
group 1 is not equal to 0
95 percent confidence interval:
 14.32900 23.90912
sample estimates:
mean in group 0 mean in group 1
        158.378         139.259
```

Exploring the associations graphically

A good picture is worth a thousand words. In addition to p-values from statistical tests, we can plot the age, sex, and maximum heart rate distributions with respect to our outcome variable. This will give us a sense of both the direction and magnitude of the relationship. First, let's plot age using a boxplot since it is a continuous variable.

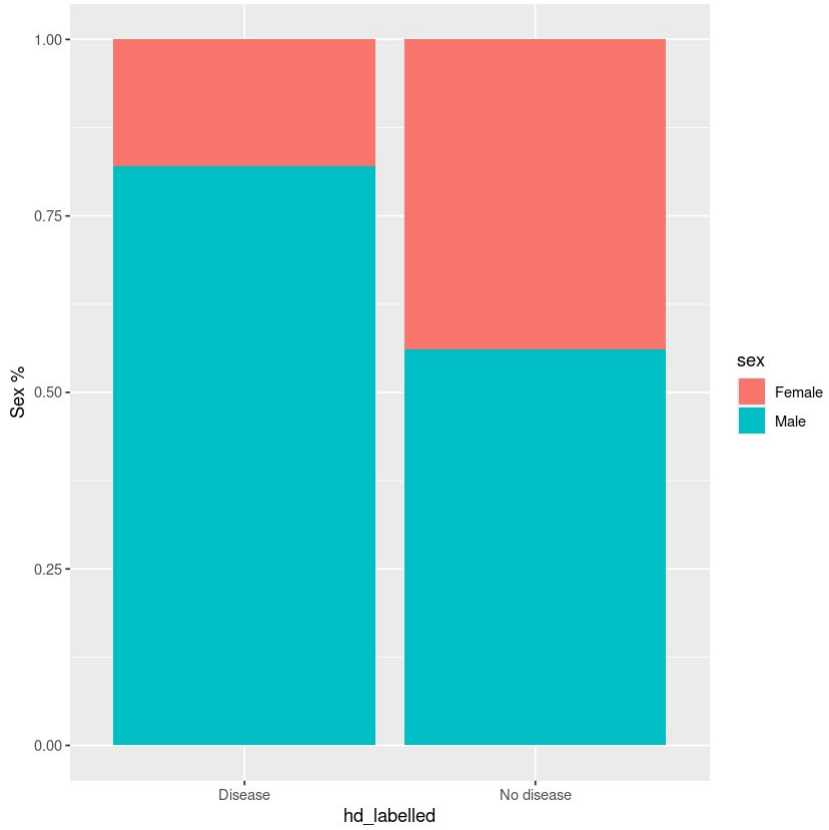
```r
# Recoding hd to be labelled directly in the dataframe
hd_data$hd_labelled <- ifelse(hd_data$hd == 0, "No disease",
"Disease")

# Plotting age vs hd using ggplot2
ggplot(data = hd_data, aes(x = hd_labelled, y = age)) + geom_boxplot()
```



Explore the associations graphically

Next, let's plot sex using a barplot since it is a binary variable in this dataset.

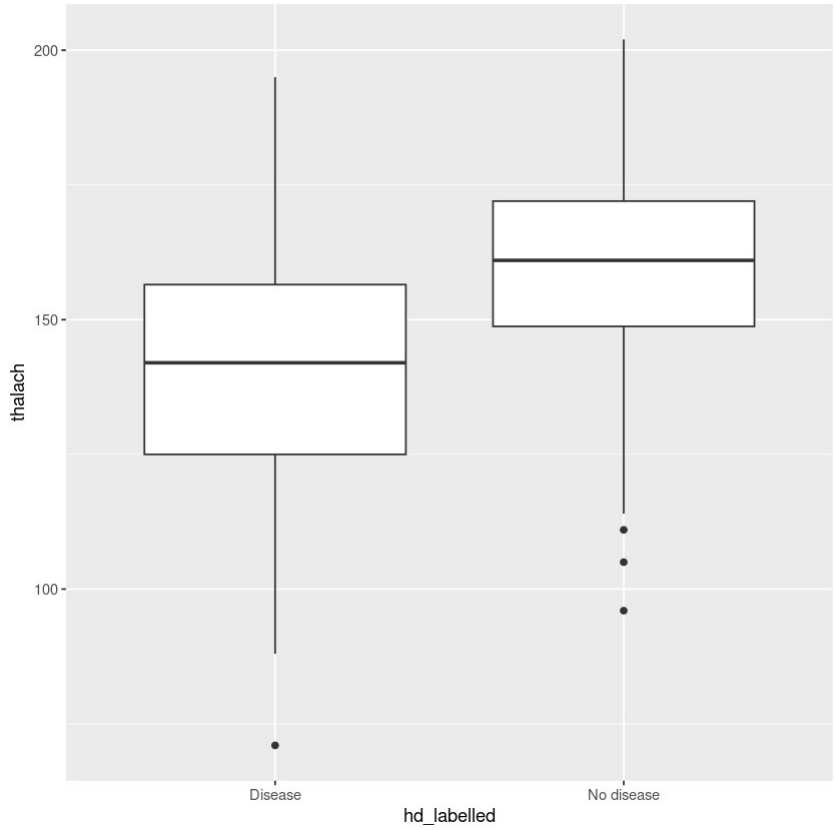
```
# sex vs hd
ggplot(data = hd_data, aes(x = hd_labelled, fill = sex)) +
geom_bar(position = "fill") + ylab("Sex %")
```



Explore the associations graphically

And finally, let's plot thalach using a boxplot since it is a continuous variable.

```
# max heart rate vs hd
ggplot(data = hd_data, aes(x = hd_labelled, y = thalach)) +
geom_boxplot()
```



Putting all three variables in one model

The plots and the statistical tests both confirmed that all the three variables are highly significantly associated with our outcome ($p<0.001$ for all tests). In general, we want to use multiple logistic regression when we have one binary outcome variable and two or more predicting variables. The binary variable is the dependent (Y) variable; we are studying the effect that the independent (X) variables have on the probability of obtaining a particular value of the dependent variable. For example, we might want to know the effect that maximum heart rate, age, and sex have on the probability that a person will have a heart disease in the next year. The model will also tell us what the remaining effect of maximum heart rate is after we control or adjust for the effects of the other two effectors. The glm() command is designed to perform generalized linear models (regressions) on binary outcome data, count data, probability data, proportion data, and many other data types. In our case, the outcome is binary following a binomial distribution.

```
# using the glm function from base R and specifying the family
argument as binomial
model <- glm(data = hd_data, hd ~ age + sex + thalach, family =
"binomial" )

# extracting the model summary
summary(model)
```

```
Call:
glm(formula = hd ~ age + sex + thalach, family = "binomial",
    data = hd_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2250  -0.8486  -0.4570   0.9043   2.1156

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.111610   1.607466   1.936   0.0529 .
age          0.031886   0.016440   1.940   0.0524 .
sexMale      1.491902   0.307193   4.857 1.19e-06 ***
thalach     -0.040541   0.007073  -5.732 9.93e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.98  on 302  degrees of freedom
Residual deviance: 332.85  on 299  degrees of freedom
AIC: 340.85

Number of Fisher Scoring iterations: 4
```

Extracting useful information from the model output

It's common practice in medical research to report Odds Ratio (OR) to quantify how strongly the presence or absence of property A is associated with the presence or absence of the outcome. When the OR is greater than 1, we say A is positively associated with outcome B (increases the Odds of having B). Otherwise, we say A is negatively associated with B (decreases the Odds of having B). The raw glm coefficient table (the 'estimate' column in the printed output) in R represents the log(Odds Ratios) of the outcome. Therefore, we need to convert the values to the original OR scale and calculate the corresponding 95% Confidence Interval (CI) of the estimated Odds Ratios when reporting results from a logistic regression.

```r
# load the broom package
library(broom)

# tidy up the coefficient table
tidy_m <- tidy(model)
tidy_m
#The model shows that being male significantly raises the outcome by
1.49 units, each year of age slightly increases it by 0.0319 units
```

```r
(borderline significant), and every unit increase in maximum heart rate during exercise decreases the outcome by 0.0405 units, with strong statistical significance.

# calculate OR
tidy_m$OR <- exp(tidy_m$estimate)

# calculate 95% CI and save as lower CI and upper CI
tidy_m$lower_CI <- exp(tidy_m$estimate - 1.96 * tidy_m$std.error)
tidy_m$upper_CI <- exp(tidy_m$estimate + 1.96 * tidy_m$std.error)

# display the updated coefficient table
tidy_m

#The model's analysis reveals that being male increases the odds of the outcome by 4.45 times (CI: 2.43 to 8.12), each additional year of age slightly raises the odds by 3% (CI: 0 to 7%), and each unit increase in maximum heart rate during exercise reduces the odds by 4% (CI: 2.6% to 5.3%), all with statistical significance.
```


```
{"table":{"data":{"estimate":[3.1116,3.19e-2,1.4919,-4.05e-2],"index":["1","2","3","4"],"p.value":[5.29e-2,5.24e-2,1.1944e-6,9.9314e-9],"statistic":[1.9357,1.9395,4.8566,-5.7319],"std.error":[1.6075,1.64e-2,0.3072,7.1e-3],"term":["(Intercept)","age","sexMale","thalach"]},"schema":{"fields":[{"name":"index","type":"string"},{"name":"term","type":"string"},{"name":"estimate","type":"float"},{"name":"std.error","type":"float"},{"name":"statistic","type":"float"},{"name":"p.value","type":"float"}],"pandas_version":"0.20.0","primaryKey":["index"]}},"total_rows":4,"truncation_type":null}
```

```
{"table":{"data":{"OR":[22.4572,1.0324,4.4455,0.9603],"estimate":[3.1116,3.19e-2,1.4919,-4.05e-2],"index":["1","2","3","4"],"lower_CI":[0.9617,0.9997,2.4347,0.947],"p.value":[5.29e-2,5.24e-2,1.1944e-6,9.9314e-9],"statistic":[1.9357,1.9395,4.8566,-5.7319],"std.error":[1.6075,1.64e-2,0.3072,7.1e-3],"term":["(Intercept)","age","sexMale","thalach"],"upper_CI":[524.3947,1.0662,8.1173,0.9737]},"schema":{"fields":[{"name":"index","type":"string"},{"name":"term","type":"string"},{"name":"estimate","type":"float"},{"name":"std.error","type":"float"},{"name":"statistic","type":"float"},{"name":"p.value","type":"float"},{"name":"OR","type":"float"},{"name":"lower_CI","type":"float"},{"name":"upper_CI","type":"float"}],"pandas_version":"0.20.0","primaryKey":["index"]}},"total_rows":4,"truncation_type":null}
```

Predicted probabilities from our model

So far, we have built a logistic regression model and examined the model coefficients/ORs. We may wonder how can we use this model we developed to predict a person's likelihood of having heart disease given his/her age, sex, and maximum heart rate. Furthermore, we'd like to translate the predicted probability into a decision rule for clinical use by defining a cutoff value on the probability scale. In practice, when an individual comes in for a health check-up, the doctor would like to know the predicted probability of heart disease, for specific values of the predictors: a 45-year-old female with a max heart rate of 150. To do that, we create a data frame called newdata, in which we include the desired values for our prediction.

```
# getting the predicted probability in our dataset using the predict()
function
# Including the argument type=”response” in order to get our
prediction.
pred_prob <- predict(model, hd_data, type="response")

# creating a decision rule using probability 0.5 as cutoff and saving
the predicted decision into the main data frame
hd_data$pred_hd <- ifelse(pred_prob >= 0.5, 1, 0)

# creating a newdata data frame to save a new case information
newdata <- data.frame(age=45, sex="Female", thalach=150)

# predicting probability for this new case and printing out the
predicted value
p_new <- predict(model, newdata, type="response")
p_new
```

```
        1
0.1773002
```

Model performance metrics

Are the predictions accurate? How well does the model fit our data? We are going to use some common metrics to evaluate the model performance. The most straightforward one is Accuracy, which is the proportion of the total number of predictions that were correct. On the other hand, we can calculate the classification error rate using 1- accuracy. However, accuracy can be misleading when the response is rare (i.e., imbalanced response). Another popular metric, Area Under the ROC curve (AUC), has the advantage that it's independent of the change in the proportion of responders. AUC ranges from 0 to 1. The closer it gets to 1 the better the model performance. Lastly, a confusion matrix is an N X N matrix, where N is the level of outcome. For the problem at hand, we have N=2, and hence we get a 2 X 2 matrix. It cross-tabulates the predicted outcome levels against the true outcome levels. After these metrics are calculated, we'll see (from the logistic regression OR table) that older age, being male and having a lower max heart rate are all risk factors for heart disease. We can also apply our model to predict the probability of having heart disease. For a 45 years old female who has a max heart rate of 150,

our model generated a heart disease probability of 0.177 indicating low risk of heart disease. Although our model has an overall accuracy of 0.71, there are cases that were misclassified as shown in the confusion matrix.

```
# load Metrics package

library(Metrics)

# calculate auc, accuracy, clasification error
auc <- auc(hd_data$hd, hd_data$pred_hd)
accuracy <- accuracy(hd_data$hd, hd_data$pred_hd)
classification_error <- ce(hd_data$hd, hd_data$pred_hd)

# print out the metrics on to screen
print(paste("AUC=", auc))
print(paste("Accuracy=", accuracy))
print(paste("Classification Error=", classification_error))

# confusion matrix
table(hd_data$hd, hd_data$pred_hd, dnn=c("True Status", "Predicted
Status")) # confusion matrix
```

```
[1] "AUC= 0.706483593612915"
[1] "Accuracy= 0.70957095709571"
[1] "Classification Error= 0.29042904290429"

           Predicted Status
True Status   0   1
          0 122  42
          1  46  93
```