STA 141A - Project Proposal: Predicting Heart Disease from Physiological Indicators Using Clinical Data

Introduction Heart disease affects a large proportion of the U.S. population and imposes significant health and economic burdens on those affected. About 27.6 million adults in the United States have been diagnosed with heart disease (Benjamin et al., 2018). According to estimates from 2003, about 50 million adults in the United States have hypertension, and about 62 million people are thought to have cardiovascular disease (Nabel, 2003). Cardiovascular disease risk is also believed to differ based on some demographic characteristics (Cooper, 2001). In order to understand risk factors associated with heart disease, we performed an analysis of one of the largest publicly available datasets for heart disease and related risk factors.

Group member contributions

- Nancy (nalhernandez@ucdavis.edu) - *Answering Research Questions, Inferences; Model setup, Editing, Function, Commenting*
- Sam (spvangorden@ucdavis.edu) - *Answering Research Questions, Inferences, Model setup, Code design*
- Robin (rkboparai@ucdavis.edu) - *Answering Research Questions, Inferences, Model setup, RMD editing, Background, Conclusion, Assist with function*

Description of the project; background of the problem and an overall goal of the project The purpose of this project is to utilize one of the largest heart disease datasets publicly available to identify possible risk factors that predict heart disease. We also hope to understand how the probability of being diagnosed given one's physiologic indicators differs according to demographic characteristics. Heart disease includes ailments that involve cardiovascular pathology (Benjamin et al., 2018).

Description of the dataset The dataset used in this analysis was obtained from the UCI Machine Learning Repository and includes data collected by the Hungarian Institute of Cardiology, University Hospitals in Zurich, Switzerland and Basel, Switzerland, the V.A. Medical Center in Long Beach, and the Cleveland Clinic Foundation. In total, the dataset includes observations on key physiologic indicators for 918 patients.

The dataset used in our analysis includes five continuous variables: age, resting blood pressure, serum cholesterol level, maximum heart rate, and depression in the ST wave induced by exercise relative to rest ("oldpeak"). The dataset includes observations for male and female adults between the ages of 28 and 77 years. Resting blood pressure (mm Hg) was collected during periods of inactivity. Serum cholesterol level (mm/dl) refers to the concentration of cholesterol in one's blood and is determined via a blood test. Maximum heart rate (beats per minute) was collected during periods of exertion. Depression in the ST segment refers to the change in the interval between ventricular depolarization and ventricular repolarization caused by exercise.

The dataset also includes seven categorical variables: chest pain type, gender, fasting blood sugar status, resting electrocardiogram (ECG) results, exercise angina status, slope of the peak exercise ST segment, and heart disease diagnosis. Chest pain type, resting ECG results, and slope of the peak exercise ST segment include more than two levels. Exercise angina status, fasting blood sugar, gender, and heart disease diagnosis are all binary variables.

```
## 'data.frame':    918 obs. of  12 variables:
##  $ Age           : int  40 49 37 48 54 39 45 54 37 48 ...
##  $ Sex           : chr  "M" "F" "M" "F" ...
##  $ ChestPainType : chr  "ATA" "NAP" "ATA" "ASY" ...
##  $ RestingBP     : int  140 160 130 138 150 120 130 110 140 120 ...
##  $ Cholesterol   : int  289 180 283 214 195 339 237 208 207 284 ...
```

```
##  $ FastingBS     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ RestingECG    : chr  "Normal" "Normal" "ST" "Normal" ...
##  $ MaxHR         : int  172 156 98 108 122 170 170 142 130 120 ...
##  $ ExerciseAngina: chr  "N" "N" "N" "Y" ...
##  $ Oldpeak       : num  0 1 0 1.5 0 0 0 0 1.5 0 ...
##  $ ST_Slope      : chr  "Up" "Flat" "Up" "Flat" ...
##  $ HeartDisease  : int  0 1 0 1 0 0 0 0 1 0 ...
```

Questions of interest

1. Do resting blood pressure, cholesterol level, and age have significant effects on one's maximum heart rate?
2. Which of these factors (*i.e.*, blood pressure, cholesterol level, and age) has the greatest influence on one's maximum heart rate?
3. Does one sex have a greater risk of heart disease compared to the other?
4. Is the relationship between heart disease and associated risk factors different across age groups?
5. Does one's cholesterol level have a significant effect on developing heart disease?
6. Does resting heart rate or blood pressure have a significant effect on developing chest pain?
7. Which variable contributes the most to predicting the presence of heart disease?
8. Given one's resting blood pressure, serum cholesterol level, and maximum heart rate, what is the probability that they have heart disease?

Methodologies

Model setup In the *model setup* phase of the analysis, we intend to use model selection techniques, including adjusted R-squared, Mallow's Cp, AIC, and/or BIC, to determine which regression models to include in our analysis. In order to identify outliers and high leverage points, we will utilize residual plots in the model setup phase as well. We also intend to assess whether our predictor variables are linearly correlated. For remedial measures, we will perform transformations to linearize a nonlinear regression relation if needed.

Answering research questions; inferences In order to answer our research questions involving continuous predictors and response variables, we intend to employ a linear regression model to determine whether continuous predictors (*e.g.*, resting blood pressure, resting heart rate, and age) significantly affect a continuous response variable (*e.g.*, serum cholesterol level). We will also use a linear model to determine whether performance on a continuous variable differs between levels of a categorical predictor (*e.g.*, sex, age group). Linear regression will be an effective approach for answering some of the research questions above because this approach entails predicting a quantitative response Y using a one or more predictors X_i.

For our research questions that involve using continuous independent variables to predict a categorical response variable (*e.g.*, being diagnosed with heart disease), we will utilize a logistic regression model. Logistic regression will be applicable in this context because "logistic regression models the *probability* that [response] Y belongs to a particular category" (Gareth et al., 2013). Logistic regression entails modeling the conditional distribution of the response variable given the predictors. In order to predict the probability that an individual has heart disease given his or her resting blood pressure, serum cholesterol level, and maximum heart rate, we can also employ the K-nearest neighbors method. KNN allows for a prediction that some observation X equals x using K training observations that are closest to x (Gareth et al., 2013).

References:

1. Benjamin, E. J., et al. (2018). Heart disease and stroke statistics—2018 Update: A report from the American Heart Association. *Circulation*, 137(12), 67-492.

2. Cooper, R. S. (2001). Social inequality, ethnicity and cardiovascular disease. *International journal of epidemiology*, 30, 48-52.

3. Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer, New York.

4. Heart Disease Data Set. *UCI Machine Learning Repository*. 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., 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

5. Nabel, E. G. (2003). Cardiovascular disease. *New England Journal of Medicine*, 349(1), 60-72.