Alex Ojemann

STAT3400 Final Project Proposal

Due 2/12/2023

What is the context of your data and why is it interesting to you?

This data is collected from five different hospitals and contains relevant indicators of cardiovascular disease. It's interesting to me because cardiovascular disease is the #1 cause of death globally and it is often said to be significantly related to factors that we can easily measure like blood pressure and cholesterol so it seems like an excellent application for regression.

Where did you get the data and why should this source be considered authoritative?

I got this data from Kaggle (https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction), but in the acknowledgements the publisher cited five reputable hospitals from which he got the data, which are authoritative because employees at these hospitals are the ones measuring and recording the data. It's useful to use this version rather than those of one of the hospitals individually because it provides a larger sample.

Is the data a random sample and what population does it represent?

This data is a random sample. It represents the populations of Budapest, Hungary, Cleveland, OH, USA, Zurich, Switzerland, Basel, Switzerland, and Long Beach, VA, USA. The Stalog (Heart) Data Set mentioned as one of the data sets combined to form this one appears to be from a combination of Zurich, Long Beach, Basel, and Budapest (https://shubamsumbria.medium.com/statlog-eda-a08e058d4f6d).

How many observations are available and what does a single observation represent?

There are 918 observations and each observation represents one person.

What is your planned response variable and what predictor variables will you consider?

The response variable is whether the person in question has cardiovascular disease, a categorical variable represented by a 0 for those that don't have cardiovascular disease and a 1 for those that do. The predictors I will use are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise induced angina, oldpeak, which refers to ST depression induced by exercise relative to rest, and ST slope, referring to the slope of the peak exercise ST segment. As mentioned in the instructions we aren't to use

measures of time as predictor variables, thus if age is considered such it will be removed.

Is each variable numerical (continuous or discrete) or categorical (nominal or ordinal)?

Age: continuous numeric

Sex: nominal categorical

Chest pain type: nominal categorical

Resting blood pressure: continuous numeric

Cholesterol: continuous numeric

Fasting blood sugar: nominal categorical (whether it's >120 mg/dl)

Resting electrocardiogram results: nominal categorical

Maximum heart rate achieved: continuous numeric

Exercise induced angina: nominal categorical

Oldpeak: Continuous numeric

ST Slope: Ordinal categorical

Heart disease: nominal categorical

Is your interest primarily inference or prediction?

My interest is prediction as the first 11 variables listed above lend themselves to predicting heart disease.