**Project: Predict Risk variables for Heart patients**

**Introduction**

Cardiovascular Disease or heart disease, generally refers to conditions that involve narrowed or blocked blood vessels that can lead to a heart attack, chest pain or stroke. Other forms of heart conditions, such as those that affect heart's muscle, valves or rhythm.

The purpose of this project is to predict the effects of different variables/parameters recorded in the data to predict patient mortality. By predicting so the physicians can determine high risk patients and can take better care of them thereby helping them survive.

**Statistical / Hypothetical Question**

Following hypothesis are tested.

* Ejection Fraction is lower for Dead Patients vs Alive Patients. Hence it is an indicator of high risk for Mortality
* Serum Creatinine is higher for Dead Patients vs Alive Patients. Hence it is an indicator of high risk for Mortality

**Outcome of EDA**

I do see some relation between some variables on the mortality.

1. Ejection Fraction - If the % age is lesser, Risk increases.

2. Creatinine Level - Wit increase in level, Risk increases.

3. Sodium Level - With decrease in level, Risk increases.

4. Age - Older the patient, higher the risk

The other factors which I observed, did not show much effect.

Iam able to support the hypothesis with the tests.

Iam able to generate prediction model (logistic regression model) and found it to be ~76% accurate with the current dataset that is being employed.

**What do you feel was missed during the analysis?**

1. Time in days between follow up visits.
2. Eating habits

While analyzing separately, I did notice time between follow up visits was making the risk lower. It is anyways evident that if you go on regular follow ups, you will be able to know the problem before it’s too late and may have the opportunity to act and decrease the risk. Eating habits also play a good role which I clearly missed in my analysis.

**Were there any variables you felt could have helped in the analysis?**

Some of the variables which I felt should have helped in analysis is the region, country and some more variables on eating habits, cholesterol levels, family history etc. which would have made the study more accurate.

**Were there any assumptions made you felt were incorrect?**

There are some concerns regarding the sample size of the data. Since the data was from a single location, there might be some other factors in play which can be due to habits of people to that specific region or part of the world and may not be reflecting in the data. If I get more geographically separated data, I might come up with better model.

**What challenges did you face, what did you not fully understand?**

I had some challenges around the Hypothesis Testing and choice of the test statistic. I chose the difference between the means of the two groups as my main test statistics, however I think I could have used some other comparisons as test statistics such as standard deviation or chi squared based tests.

**References:**

<https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118#:~:text=The%20term%20%22heart%20disease%22%20is,pain%20(angina)%20or%20stroke>.

<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>

<https://www.healthline.com/health/creatinine-blood#results>

Discover Statistics Using R, Andy Field | Jeremy Miles | Zoe Field

R for Everyone, Jared P Lander

Think Stats, Allen B Downey