STA141A Final

STA141A | Fall 2022 | Kühnert

Alvin Akpokli, Sarvesh Krishan, Ngai Pan Ng, Hannah Spray

7 December 2022

Project Overview

This project aims to investigate possible risk factors for death due to heart failure.

Group 4 members:

- Alvin Akpokli (akakpokli@ucdavis.edu)
- Sarvesh Krishan (skrishan@ucdavis.edu)
- Ngai Pan Ng (npng@ucdavis.edu)
- Hannah Spray (hgspray@ucdavis.edu)

Research questions of interest:

1. Is death from heart failure significantly more prevalent in those with high blood pressure (i.e. hypertension)?
2. Is death from heart failure significantly more prevalent in men or women that smoke versus those who do not?

Dataset:

Our dataset is called "Heart failure clinical records dataset". This dataset was obtained from the UCI Machine Learning Repository. This dataset contains both boolean and numerical variables:

- age: age of the patient (years)
- anaemia: decrease of red blood cells or hemoglobin (boolean)
- high blood pressure: if the patient has hypertension (boolean)
- creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L)
- diabetes: if the patient has diabetes (boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- platelets: platelets in the blood (kiloplatelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)

- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (days)
- [target] death event: if the patient deceased during the follow-up period (boolean)

We will treat "death event" as our binary response variable.

Method:

We plan to conduct our analyses as a classification problem, using a logistic regression model.

Diagnostics

Logistic Regression Model Assumptions

We can begin our analysis by verifying that our dataset meets the model assumptions for logistic regression:
1. dependent variable is binary
2. observations are independent
3. little to no multicollinearity between predictor variables

We already know that our dependent variable, 'DEATH_EVENT' is binary - it will equal 1 if the given patient deceased during the follow-up period, and 0 otherwise. Regarding independence between observations, we can assume that this assumption is met through the study design - i.e. the sampling procedure used for this dataset utilized independent random sampling (<- humus just wrote this part and is not sure if this is actually true).

To check the assumption regarding multicollinearity between predictor variables, we can utilize the variance inflation factor (VIF) to detect multicollinearity in our model. We begin by creating our "full model", regressing our response variable (DEATH_EVENT) with all of the predictor variables:

```
full_model = glm(DEATH_EVENT ~ ., family = "binomial", data = data)
```

And investigate the VIF for each predictor variable:

Table 1: Table 1: VIF for predictors

	VIF
age	1.1043
anaemia	1.1145
creatinine_phosphokinase	1.0856
diabetes	1.0524
ejection_fraction	1.1728
high_blood_pressure	1.0630
platelets	1.0453
serum_creatinine	1.1021
serum_sodium	1.0707
sex	1.3806
smoking	1.2845
time	1.1518

Model Fitting

Discussion

References

Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16 (2020).