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 Y, where

$$Y = \begin{cases} 1 & \text{if patient } i \text{ deceased during the follow-up period} \\ 0 & \text{otherwise} \end{cases}$$

Method:

We plan to conduct our analyses as a classification problem, using a logistic regression model. We will create a model that expresses the log-odds of a patient being deceased during the follow-up period as a function of our predictor variables. Our goals for these fitted models are:
1. To determine whether or not high blood pressure (hypertension) increases the probability of a patient being deceased during the follow-up period
2. To determine whether or not _____ blaha lhah?

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)
```

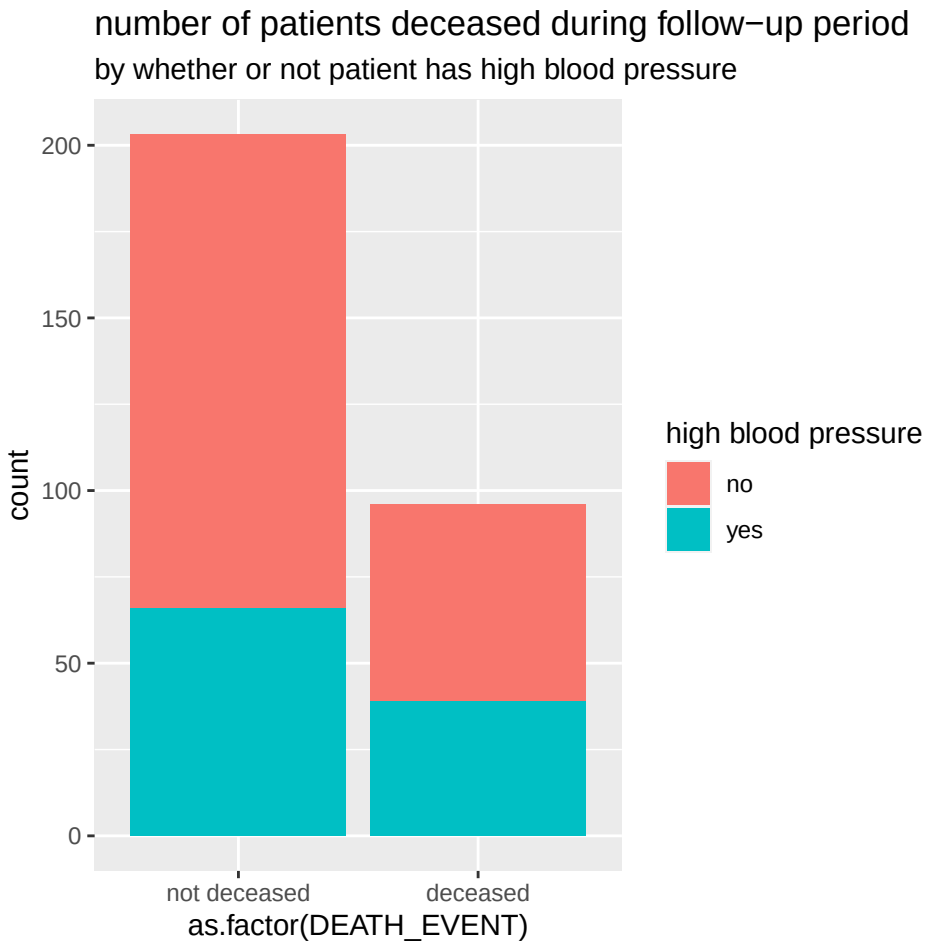
And investigate the VIF for each predictor variable:

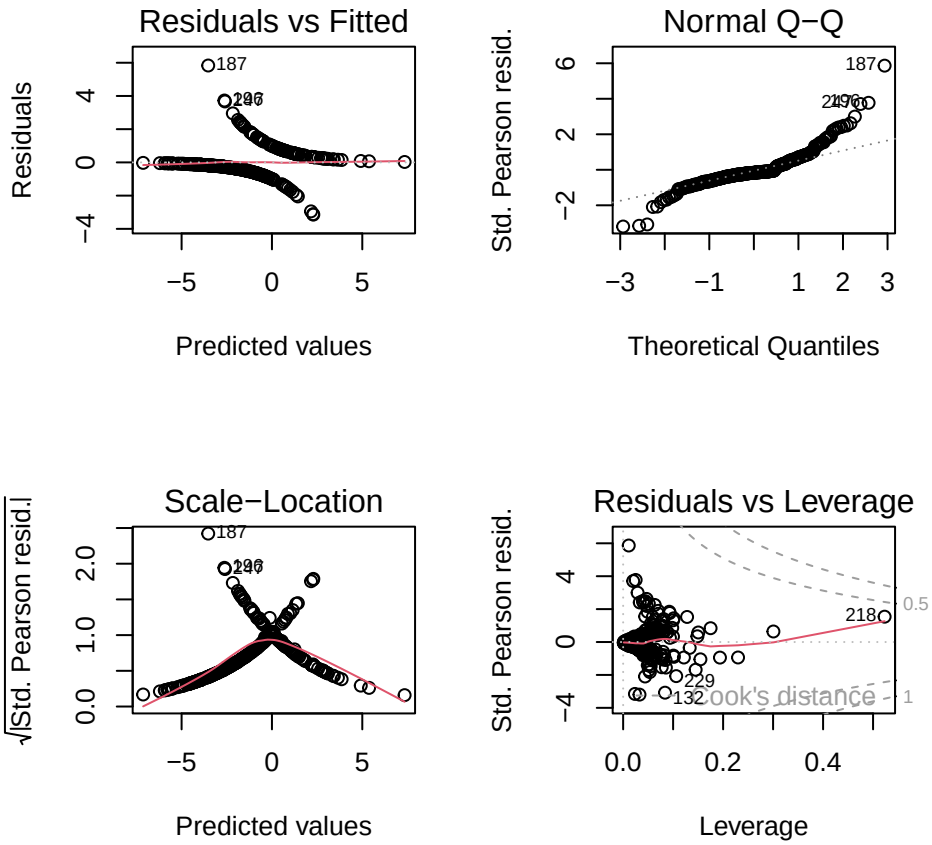
Table 1: Table 1: VIF for predictors

	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

For the purposes of this analysis, we can use the general rule of thumb which states that predictor variables with VIF values greater than 5 should be treated as problematic. We can see by looking at the table above that none of the predictor variables in the fitted model have VIF values greater than 5; in fact, all of the values are between 1 and 2. Hence, we can proceed without having to remedy multicollinearity.

some plots:


number of patients deceased during follow−up period
by whether or not patient has high blood pressure
200
150
100
50
0
count
not deceased
deceased
as.factor(DEATH_EVENT)
high blood pressure
no
yes



Model Fitting

Principle Component Analysis

Table 2: mean squared prediction error for 10 different pcr models

number of principle components	mean squared prediction error
1	0.38474
2	0.29188
3	0.25216
4	0.27941
5	0.28317
6	0.26459
7	0.22429
8	0.18658
9	0.23487
10	0.25861

hannah note to self:
- how to interpret mean squared prediction error?
- how to know how many pcr models to test? like for the example above i computed 10 different RMSE's

but how do we decide how many RMSEs we want to compare?
- is comparing RMSEs the right way to go about this? im copying hw6 not sure what to do see hw6

Hypothesis Testing

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).