Heart Failure Prediction

Research scenario and questions:

Nowadays, cardiovascular diseases (CDVs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, accounting for 31% of all deaths worldwide. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as smoking, unhealthy diet and obesity, physical inactivity, and harmful alcohol use via population-wide strategies. Prediction of these risk factors by an appropriate model using the clinical data would help reduce the mortality of heart failure via lifestyle adjustment or medicinal intervention. There are several questions we want to answer by the study. First, is there a difference between heart failure-caused death among men versus women? Second, whether the factors in the clinical records dataset are associated with heart failure at $\alpha = 0.05$ level. Third, if so, what are the factors and the corresponding odds ratio?

Discerption of data set

The dataset is downloaded from Kaggle (https://www.kaggle.com/andrewmvd/heart-failure-clinical-data). It has 13 columns in total, in which anemia, diabetes, high blood pressure, sex (0 represents women, 1 represents men), smoking, and death event are dichotomous variables, age, creatinine phosphokinase (mcg/L), ejection fraction (percentage), platelets (kiloplatelets/mL), serum creatinine (mg/dL), serum sodium (mEq/L), and time (follow-up period, day) are continuous variables. The data can be used for analysis without any cleaning. We are interested in the association between heart failure-caused death and age, anemia, creatinine phosphokinase level, diabetes, ejection fraction, high blood pressure, platelets number, serum creatinine level, serum sodium level, sex, and smoking.

Statistic methods

A two-sample test for proportions is performed to test if the heart failure-caused death is the same across women and men. Since the response variable is dichotomous, multiple logistic regression with death event as the response variable and the other 11 variables as the explanatory variables was performed. First, a global test (Wald test) was performed to test whether or not any associations exist between the explanatory variables and the response variable. Then, evaluating tests based on each explanatory variable (t-test) were performed. The ROV curve for the multiple logistic regression model was applied to check how well the model predicts the death caused by heart failure.

Data analysis and results

Table 1. Summary of the heat failure-caused death by sex

Population	Population description	Sample size	Count of success	Count of failure	Sample proportion
0	Women	203	132	71	0.650
1	Men	96	62	34	0.646

Significance tests for differences in proportions

H_0: $p_1=p_2$ (the underlying population proportion died from heart failure among women is the same as the population portion died from heart failure among men)

H_1: $p_1 \neq p_2$ (the underlying population proportion died from heart failure among men is different from the population portion died from heart failure among women)

Since the p-value for the two-samples proportions test is bigger than alpha ($p = 1$, $\alpha = 0.05$), we failed to reject the null hypothesis. There is no difference between the population proportion who died from heart failure among men versus women.

Global test for the multiple logistic regression

H_0: $\beta_{age} = \ldots\ldots = \beta_{smoking} = 0$ (there is no association between death caused by heart failure and any of the variables)

H_1: there is at least one $\beta_i \neq 0$ (at least one variable is associated with the death caused by heart failure)

The global test results (Chi-squared test, p = 3e-14) of the multiple logistic regression model showed that at least one of the 11 explanatory variables is associated with the event of death caused by heart failure.

Test of each explanatory variable

Summary of the tests:

Coefficients:

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.498e+00	7.942e-01	1.886	0.0603 .
age	9.012e-03	2.097e-03	4.298	2.36e-05 ***
anaemia	5.488e-02	5.023e-02	1.093	0.2755
creatinine_				

phosphokinase	4.880e-05	2.561e-05	1.905	0.0577 .
diabetes	1.664e-02	5.038e-02	0.330	0.7415
ejection_fraction	-1.059e-02	2.105e-03	-5.032	8.57e-07 ***
high_blood_pressure	6.795e-02	5.120e-02	1.327	0.1855
platelets	-7.211e-08	2.523e-07	-0.286	0.7752
serum_creatinine	1.063e-01	2.412e-02	4.409	1.47e-05 ***
serum_sodium	-1.096e-02	5.726e-03	-1.914	0.0566 .
sex	-6.256e-02	5.836e-02	-1.072	0.2847
smoking	1.340e-02	5.844e-02	0.229	0.8187

Age:

H_0: $\beta_{age} = 0$ or $OR_{age} = 1$ (there is an association between age and risk of death caused by heart failure)

H_1: $\beta_{age} \neq 0$ or $OR_{age} \neq 1$ (there is no association between age and risk of death caused by heart failure)

We reject the null hypotheses after adjusting for other variables since p = 2.36e-05 < α. We have significant evidence at the $\alpha = 0.05$ level that $\beta_{age} \neq 0$. That is, there is evidence of an association between age and the risk of death caused by heart failure after adjusting for other variables. The odds ratio for death caused by heart failure is $e^{\beta age} = 1.009$ for every year increase in age after adjusting for other variables.

Ejection fraction:

H_0: $\beta_{ejection\_fraction} = 0$ or $OR_{ejection\_fraction} = 1$ (there is an association between ejection fraction and risk of death caused by heart failure)

H_1: $\beta_{ejection\_fraction} \neq 0$ or $OR_{ejection\_fraction} \neq 1$ (there is no association between ejection fraction and risk of death caused by heart failure)

We reject the null hypotheses after adjusting for other variables since p =8.57e-07 < α. We have significant evidence at the $\alpha = 0.05$ level that $\beta_{ejection\_fraction} \neq 0$. There is evidence of an association between ejection fraction and the risk of death caused by heart failure after adjusting for other variables. The odds ratio for death caused by heart failure is $e^{\beta ejection\_fraction} = 0.989$ for every 1percent increase in ejection fraction after adjusting for other variables.

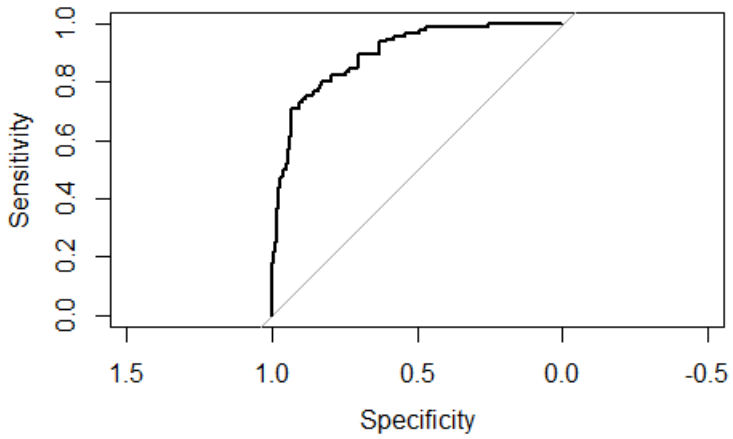
Serum creatinine:

> H_0: $\beta_{serum\_creatinine} = 0$ or $OR_{serum\_creatinine} = 1$ (there is an association between serum creatinine level and risk of death caused by heart failure)
>
> H_1: $\beta_{serum\_creatinine} \neq 0$ or $OR_{serum\_creatinine} \neq 1$ (there is no association between serum creatinine level and risk of death caused by heart failure)
>
> We reject the null hypotheses after adjusting for other variables since p =1.47e-05 < α. We have significant evidence at the α = 0.05 level that $\beta_{serum\_creatinine} \neq 0$. There is evidence of an association between serum creatinine level and the risk of death caused by heart failure after adjusting for other variables. After adjusting for other variables, the odds ratio for death caused by heart failure is eβserum_creatinine = 1.112 for every 1 mg/dL increase in serum creatinine level.

We do not have sufficient evidence at the α = 0.05 level that there is an association between death caused by heart failure and any of the variables, anemia, creatinine phosphokinase, diabetes, high blood pressure, platelets, serum sodium, sex, and smoking, respectively.

C-statistic and ROC curve for the model

The c-statistic for the model is 0.807, indicating the model has a good prediction for death caused by heart failure. Below is the ROC curve for the model.



Conclusion and discussion

In a nutshell, there is no difference in the risk of death caused by heart failure between men and women. Among the listed 11 factors recorded in the data, age, ejection fraction, and serum creatinine level are predictive of the event of death caused by heart failure after adjusting for other factors. The odds ratio for death caused by heart failure increases 0.9% for every year increase in

age, decreases 1.1% for every 1% increase in ejection fraction, and increases 11.2% for every 1 mg/dL increase in serum creatinine level. The c-statistic indicates that this is a pretty strong model for predicting the death caused by heart failure using the 11 factors.

To our surprise, some factors commonly known as significant associations with heart failure, such as high blood pressure, serum sodium level, are not predictive in this model. This might be caused by inappropriate data processing and data selection. For example, rather than being continuous, the variable high blood pressure is dichotomous. As normal pressure ranges vary by age and sex, simply dividing the blood pressure into high blood pressure groups and normal groups is arbitrary. Moreover, collinearity may exist between the high blood pressure and serum sodium level, leading to unreliable regression coefficients. Thus, presenting the data in an appropriate data type and checking the possible collinearity between variables before analysis could improve the reliability and stability of the model.