Kadir Orçun Altunel

Koa94097@huskies.bloomu.edu

THE EFFECT OF SERUM CREATITINE AND EJECTION FRACTION ON HEART FAILURE

Table of Contents

**Abstract……………………………………………………………………………1**

Background…………………………………………………………………1

Methods……………………………………………………………………..1

Results………………………………………………………………………1

Conclusion…………………………………………………………………..2

**Background………………………………………………………………………..2**

Dataset……………………………………………………………………….5

**Methods……………………………………………………………………………7**

Logistic Regression………………………………………………………….7

AIC…………………………………………………………………………..8

Chi-Square Test……………………………………………………………...9

Association of Predicted Probabilities and Observed Responses………….10

Confusion Matrix and Statistics……………………………………………12

ROC And AUC…………………………………………………………….13

**Results and Conclusions…………………………………………………………20**

**References………………………………………………………………………..23**

**Abstract**

*Background*

Cardiovascular diseases are one of the leading causes of death globally. Just in 2019, an estimated 17.9 million people died from cardiovascular diseases which accounts for 32% of all global deaths. Cardiovascular diseases are conditions affect the structure or the function of the heart. Heart failure is one of the most prevalent cardiovascular diseases. It occurs when the heart doesn’t pump enough blood to the body resulting in the body to hold in salt and water, which in return causes swelling and shortness of breath. It is projected that the number of people diagnosed with heart failure will increase by 46% in 2030, which leaves 8 million patients with heart failure according to the American Heart Association. Since heart is one of the most vital organs, predicting heart failure and finding risk factors that may cause heart failure has become a priority for medical doctors and physicians.

*Methods*

In this paper, I analyzed a dataset of 299 patients with heart failure collected in 2015. I applied logistic regression to both predict the survival of patients and rank the features corresponding to the most important risk factors. Logistic regression is an extremely powerful algorithm, even for very complex problems, it does a spectacular job. Linear Regression is good not only for prediction; once you have a fitted Linear Regression model you can also learn things about relationships between the response and the predictive variables. All these characteristics of Logistic Regression makes it an indispensable tool for this paper.

*Results*

My results demonstrated that serum creatinine, ejection fraction and age play a significant role in the survival of patients with heart failure. Using these 3 predictors alone can lead to more accurate predictions regarding the chances of patients’ survival.

*Conclusions*

The result of this analysis can have serious consequences in saving patients’ lives. It can help doctors create tools that may guide them to estimate the chance of survival rate for patients with heart failure and inspire scientists to investigate the effects of serum creatinine and ejection fraction on the survival from heart failure.

**Background**

Cardiovascular diseases are conditions that affect the structures or function of the heart. It is the leading cause of death in the U.S. About 659,000 people in the U.S. die from heart disease each year; that’s 1 in every 4 deaths. Most cardiovascular disease affects older adults. In the United States 11% of people between 20 and 40 have CVD, while 37% between 40 and 60, 71% of people between 60 and 80, and 85% of people over 80 have cardiovascular diseases. Heart failure is a condition that develops when heart doesn’t pump enough blood for body’s needs. This can happen if heart can’t fill up with enough blood or it is too weak to pump properly. The term Heart failure is a serious condition that needs medical care. Three main phenotypes describe heart failure according to the measurement of the left ventricle ejection fraction (EF), and the differentiation between these types is important due to different demographics, co-morbidities, and response to therapies:

* Heart failure with reduced ejection fraction (HFrEF): EF less than or equal to 40%
* Heart failure with preserved EF (HFpEF): EF is greater than or equal to 50%
* Heart failure with mid-range EF (HFmrEF) (other names are: HFpEF-borderline and HFpEF-improved when EF in HFrEF improves to greater than 40%): EF is 41% to 49% per European guidelines and 40 to 49% per the US guidelines. A new class of HF that introduced by the 2016 European Society of Cardiology (ESC) guidelines for diagnosis and management of HF. This class was known as the grey area between the HFpEF and HFrEF and now has its distinct entity by giving it a name as HFmrEF.

All patients with HFrEF have concomitant diastolic dysfunction; in contrast, diastolic dysfunction may occur in the absence of systolic dysfunction.

Doctors usually classify patients' heart failure according to the severity of their symptoms. The table below describes the most used classification system, the New York Heart Association (NYHA) Functional Classification. It places patients in one of four categories based on how much they are limited during physical activity.

Text

Description automatically generated

Given the importance of a vital organ like heart, predicting cardiovascular diseases has become a priority for doctors. Though there have been various predictions made before, most of them failed to provide high accuracy. However, electronic health records (EHRs) have been a crucial tool to provide a great source of information regarding hidden and not so obvious correlations between patients’ data and risk factors to assess and better understand cardiovascular disease risks.

In this paper, I analyzed a dataset of medical records of patients having heart failure released by Ahmad and colleagues in July 2017. Ahmad and colleagues employed traditional biostatistics time-dependent models to predict mortality and identify the key features of 299 Pakistan patients having heart failure, from their medical records.

I used logistic regression to predict survival rate of patients with heart failure. Using the dataset, I implemented various methods such as backward stepwise regression and Area Under the Curve” (AUC) of “Receiver Characteristic Operator” (ROC) which will be discussed in *Methods* section of the paper. Finally, I discuss the results from the methodology and draw conclusions in *Results* section of the paper.

***Dataset***

I analyzed a dataset containing the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All 299 patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure.

Graphical user interface, text, application

Description automatically generated

The dataset contains 13 features, which report clinical, body, and lifestyle information. Some features are binary: anemia, high blood pressure, diabetes, sex, and smoking.

Creatine phosphokinase (CPK) is an enzyme in the body. It is found mainly in the heart, brain, and skeletal muscle. When the total CPK level is very high, it most often means there has been injury or stress to muscle tissue, the heart, or the brain.

Ejection fraction (EF) refers to how well left ventricle (or right ventricle) pumps blood with each heartbeat. Most times, EF refers to the amount of blood being pumped out of the left ventricle each time it contracts. The left ventricle is the heart's main pumping chamber. EF is expressed as a percentage. An EF that is below normal can be a sign of heart failure. A normal left ventricular ejection fraction (LVEF) range from 55% to 70%. An LVEF of 65%, for example, means that 65% of the total amount of blood in the left ventricle is pumped out with each heartbeat.

Platelets, also called thrombocytes, are small blood cells that help blood to clot. When a blood vessel gets damaged, platelets gather at the damaged site and make a plug (clot). Clotting helps slow down and stop bleeding and helps wounds heal.

Serum creatinine test is a measure of how well kidneys are performing their job of filtering waste from blood.

Creatinine is a chemical compound left over from energy-producing processes in your muscles. Healthy kidneys filter creatinine out of the blood. Creatinine exits body as a waste product in urine. An increased level of creatinine may be a sign of poor kidney function.

Serum Sodium test measures the amount of sodium in blood. Sodium is a type of electrolyte. Electrolytes are electrically charged minerals that help maintain fluid levels and the balance of chemicals in body called acids and bases. Sodium also helps nerves and muscles work properly. If serum sodium levels are too high or too low, it may mean that there might be a problem with kidneys, dehydration, or another medical condition.

**Methods**

In this section, I listed the methods I used in logistic regression for binary classification of survival. Using the methods listed below, I aimed to pinpoint which predictors were highly significant in the survival detection of heart failure.

***Logistic Regression***

Logistic regression, also called a logit model, is used to model dichotomous outcome variables. In the logit model the log odds of the outcome are modeled as a linear combination of the predictor variables. The model is written as:

Company name

Description automatically generated with low confidence

Fitting my full model on R looked very similar to fitting a simple linear regression. Instead of lm () we use glm (). The only other difference is the use of family = "binomial" which indicates that we have a two-class categorical response.

Here are the results of my full model:

Text, table

Description automatically generated with medium confidence

Looking at the R output, Ejection Fraction, Serum Creatinine and Time are highly significant since their p values are low. It is also seen that Age and Serum Sodium can be considered significant in the full model.

***AIC***

The Akaike Information Criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

* the number of independent variables used to build the model.
* the maximum likelihood estimates of the model (how well the model reproduces the data).

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables.

To compare models using AIC, you need to calculate the AIC of each model. If a model is more than 2 AIC units lower than another, then it is considered significantly better than that model. Later, we will see if we can make our model significantly better by comparing with other models.

The formula for AIC is written as:

Text, whiteboard

Description automatically generated

***Chi-Square Test***

Chi-square formula is used to compare two or more statistical data sets. The chi-square formula is used in data that consist of variables distributed across various categories and helps us to understand whether that distribution is different from what one would expect by chance. It’s extremely useful to test goodness of fit in linear regression by investigating Analysis of Deviance Table. The formula is written as:

A picture containing text

Description automatically generated

Running Chi-Square Test on the full model, R output gives the results below:

A screenshot of a computer

Description automatically generated with medium confidence

The difference between the null deviance and the residual deviance is a good indicator of how the full model is doing against the null model (a model with only the intercept). The wider the gap is, the better the fit is. Analyzing the table, we can see the drop in deviance when adding each variable one at a time.

***Association of Predicted Probabilities and Observed Responses***

This association is an important procedure assessing the predictive ability of a model. They are based on the number of pairs of observations with different response values, the number of concordant pairs, and the number of discordant pairs. A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score

than the observation with the higher ordered response value. If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant. If the pair is neither concordant nor discordant, it is a tie.

* **Somer’s D**

Somers’ D which is short for Somers’ Delta, is a measure of the strength and direction of the association between an ordinal dependent variable and an ordinal independent variable. Somer’s D can be a good indicator of fitness of model in logistic regression.

An ordinal variable is one in which the values have a natural order (e.g., “bad”, “neutral”, “good”). The value for Somers’ D ranges between -1 and 1 where:

-1: Indicates that all pairs of the variables disagree

1: Indicates that all pairs of the variables agree

The formula is written as:



* **Goodman-Kruskal Gamma**

Goodman-Kruskal gamma (γ) shows how many more concordant than discordant pairs exist divided by the total number of pairs excluding ties. The use the Goodman-Kruskal gamma to measure the association between the ordinal variables is another important indicator of the fitness of model in logistic regression.

Perfect association exists when |γ| = 1. In ordinal and binary logistic regression, if X and Y are independent, then γ = 0.

The formula is written as:

**GAMMA =** **(NC – ND) / (NC + ND)**

* **Kendall’s Tau**

Another indicator of the fitness of the model in logistic regression Kendall’s Tau, which measures the relationship between two columns of ranked data. Again, the value of a correlation coefficient can range from -1 to 1, with -1 indicating a perfect negative relationship, 0 indicating no relationship, and 1 indicating a perfect positive relationship.

**TAU = (C-D) / (C+D)**

Looking at the R output of our full model:

Table

Description automatically generated

We see that Somer’s D, Gamma and Concordance correlation coefficients are relatively high which means our model is highly significant.

***Confusion Matrix and Statistics***

One common way to evaluate the quality of a logistic regression model is to create a confusion matrix, which is a 2×2 table that shows the predicted values from the model vs. the actual values from the test dataset.

Confusion matrix can indicate important information regarding the fit of the model.

**Sensitivity:** The “true positive rate” – the percentage of individuals the model correctly predicted would default.

**Specificity:** The “true negative rate” – the percentage of individuals the model correctly predicted would not default.

Looking at our R output:

A screenshot of a computer

Description automatically generated with low confidence

We see that the accuracy of our model is very high at 85%, misclassification error rate is at 15% and specificity of our model is at 92%. These results indicate that our model is a good fit.

***ROC And AUC***

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model can distinguish between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. By analogy, the Higher the AUC, the better the model is.

Here is our ROC:

Chart

Description automatically generated

And our AUC is:



This result shows that our model is good standing at almost 0.9.

Now that we know more about our full model, we will go ahead and look for a better model with lower AIC and compare the results we get from the optimal model to the full model.

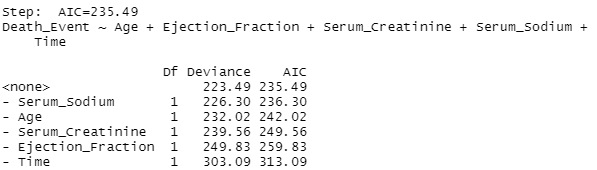
We will use stepwise regression approach which at each step gradually eliminates variables from the regression model to find a reduced model that best explains the data.

Here’s our R Output that shows us the reduced model that best explains the data:

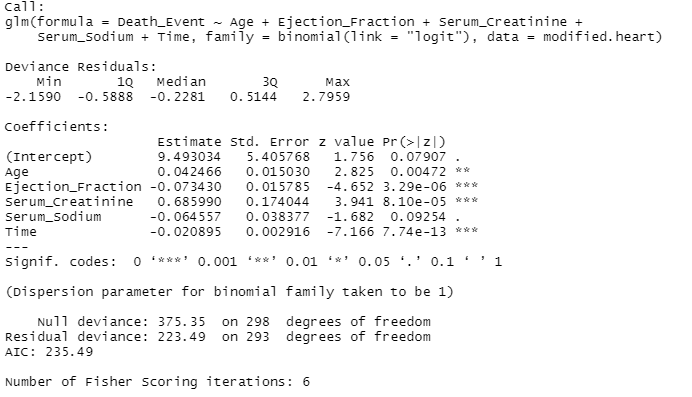
Chart, line chart

Description automatically generated

Our stepwise regression approach shows us that using variables Time, Ejection Fraction, Serum Creatinine, Age and Serum Sodium significantly improves our model with AIC at 235.49.



We go ahead and look at the summary of our reduced model on R:



We see that our model is highly significant with all the predictors in the model suggest that the reduced model is also a good fit improved AIC.

A screenshot of a computer

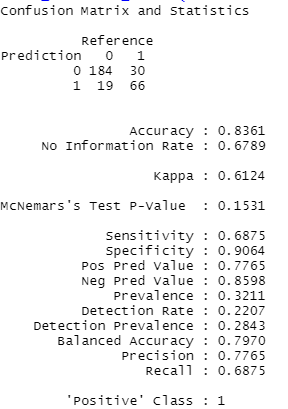
Description automatically generated with medium confidence

Again, the difference between the null deviance and the residual deviance is a good indicator of how the full model is doing against the null model (a model with only the intercept). The wider the gap is, the better the fit is. Analyzing the table, we can see the drop in deviance when adding each variable one at a time.

Table

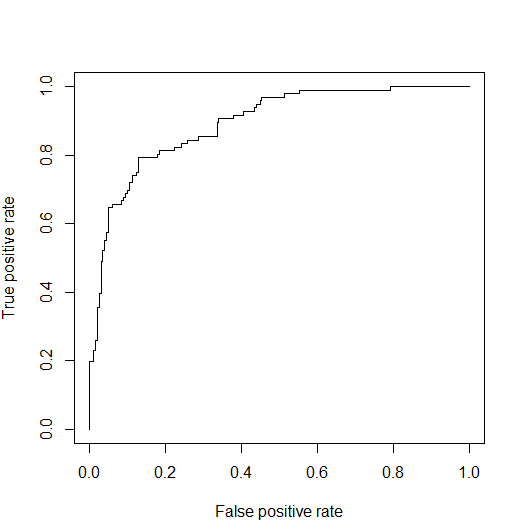
Description automatically generated

Somers’ D, Gamma, Tau-a and c correlation coefficients are highly significant in this model as well.



Our confusion matrix and statistics also show a high accuracy rate at 0.84 with specificity rate at 0.91

Here’s our ROC:



And our AUC is:



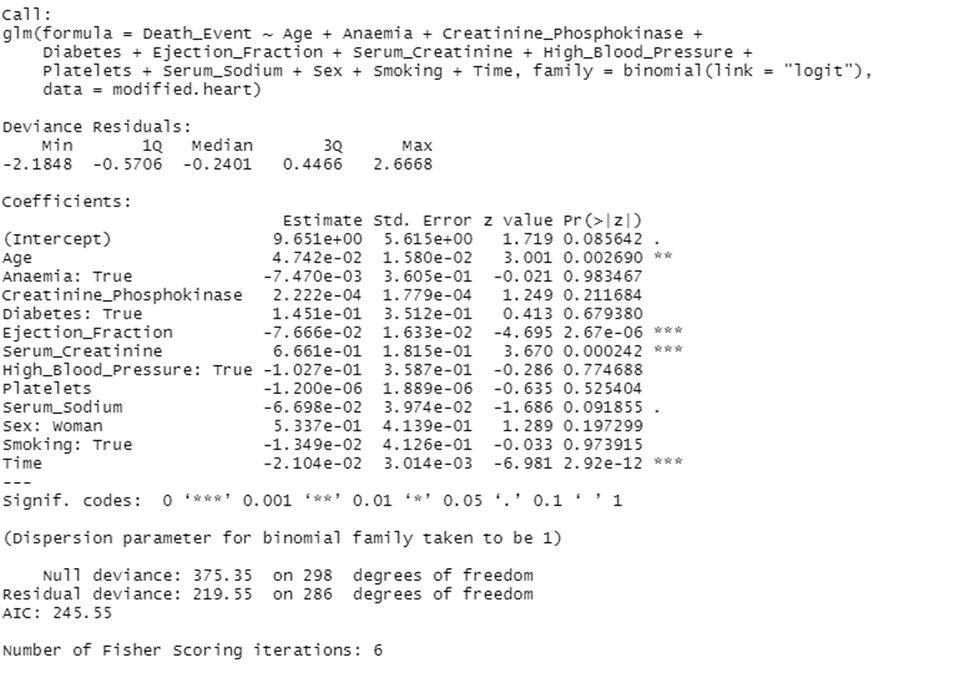
This result also shows that our model is a good fit at 0.89

**Results and Conclusions**

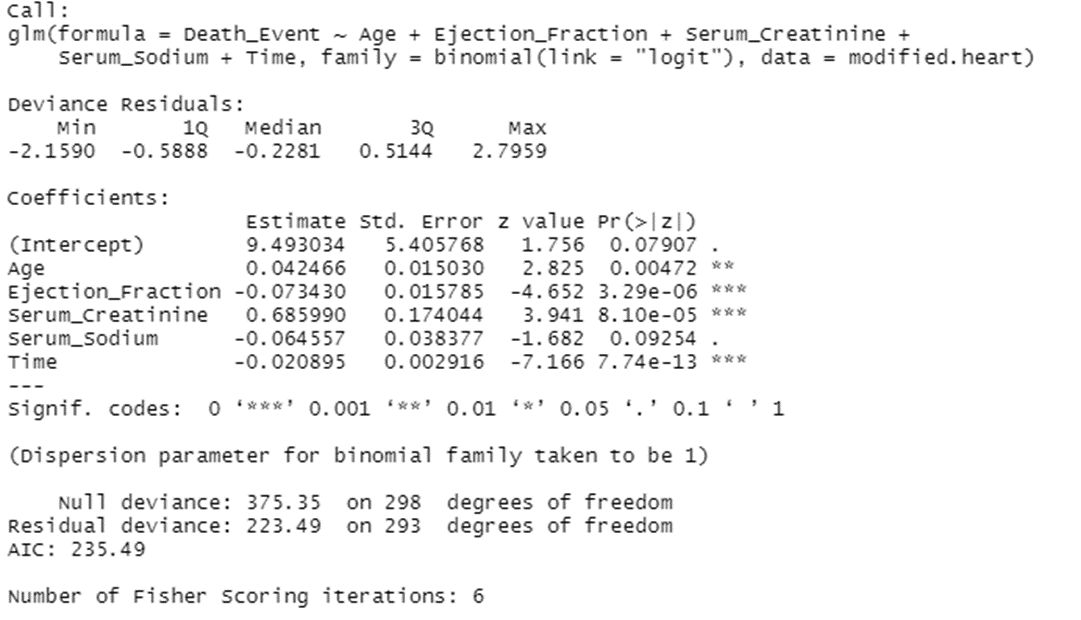
Comparing our full model and reduced model, we can clearly see that using significant predictors alone can help doctors predict heart failure survival in patients.

The results from our full model and reduced model didn’t differ as much except the fact we have seen an improved AIC on reduced model which made our prediction more accurate.

Full model:



Reduced model:



As we can see, our AIC went down by 10 points which is a great indication of a better fit for our model.

A screenshot of a computer

Description automatically generated with low confidence

Moreover, looking at the importance of the predictors, we see that our reduced model has the predictors that has the importance to our prediction. We see that Age, Ejection Fraction, Serum Creatinine and Time are highly significant to the model which are all presented in our reduced model.

Consequently, we can deduce that doctors can determine the probability of death due to heart failure by just looking at ejection fraction and serum creatinine (based on bloodwork). These two blood indices are highly important in patients with heart failure. These indices can help doctors save time in their treatment without digging into other markers on blood work.

As a limitation of the study, I must report the small size of the dataset. Larger dataset would have permitted me to obtain more reliable results. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases.

**References:**

<https://stats.idre.ucla.edu/r/dae/logit-regression/>

<https://support.sas.com/documentation/onlinedoc/stat/143/logistic.pdf>

<https://support.minitab.com/en-us/minitab/19/help-and-how-to/statistics/tables/supporting-topics/other-statistics-and-tests/what-are-the-goodman-kruskal-statistics/>

<https://cran.r-project.org/web/packages/blorr/vignettes/introduction.html>

<https://www.r-bloggers.com/2015/09/how-to-perform-a-logistic-regression-in-r/>

<https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

<https://towardsdatascience.com/introduction-to-aic-akaike-information-criterion-9c9ba1c96ced>

<https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5>