Logistic Regression: Death following Heart Failure

Philip Loewen, Marisa Ortiz, Ci Xu, Rohan Joseph

2023-04-13

Introduction

The threat of heart disease and heart failure deeply permeates modern society. According to the World Health Organization, heart disease causes 31% of deaths globally. In Canada, about 1 in every 12 adults over the age of 20 lives with heart disease. In Pakistan, where our data set is sourced, this number is much higher at 1 in every 4 middle-aged Pakistanis suffering from heart disease. In fact, this illness has seen an increase in prevalence in Pakistan in recent years and is now considered an epidemic by many medical professionals. Hospital readmissions are unfortunately common and represent a significant burden on the individual, the community, and the healthcare system. Patient education and awareness are a critical part of heart failure care as they can provide patients with more agency and effectiveness in self-management.

For our final project in STAT 306, synthesizing what we've learned about finding relationships in data, we've decided to explore the "heart failure clinical records" data set. This data set contains the medical records of 299 heart failure patients admitted to the Institute of Cardiology and Allied Hospital in Faisalabad, Pakistan, between April and December 2015. Thirteen clinical features were collected during the patients' follow-up period: age, anaemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, platelets, sex, serum creatinine, serum sodium, smoking, length of follow-up period, and death event (if the patient deceased during the follow-up period), with survival of the patient taking a death event value of 0 and death taking value 1. The motivation behind this analysis was to model patient mortality as a function of these clinical features and investigate the major risk factors associated with heart disease.

The aim of our project is to iteratively build multivariate logistic models with the data at hand to arrive at the most efficient relationship to predict a death event following heart failure. Our research question is: ***Which variables create the most efficient model to predict a death event in patients who have experienced heart failure?***

The identification of these variables can help narrow the focus of clinical care by allowing physicians to predict the health and mortality of patients based on fewer clinical features. Creating more concise patient profiles to address the health outcomes of heart failure patients can alleviate the social and structural burdens heart disease has on communities around the world by allowing quicker and cheaper diagnoses as less tests must be done to predict death.

Analysis

Data Exploration

Firstly, we will visualize the data. Visualizing the data allows us to preemptively search for potential issues that we may encounter in the future. One common issue is not having representative data, as such we visualize the age distribution and the sex proportions in figure 1. We notice that there are quite a few more males than females in this data set. Although this is not necessarily a problem, we may want to be careful when creating our model as including sex as a variable might require attention.

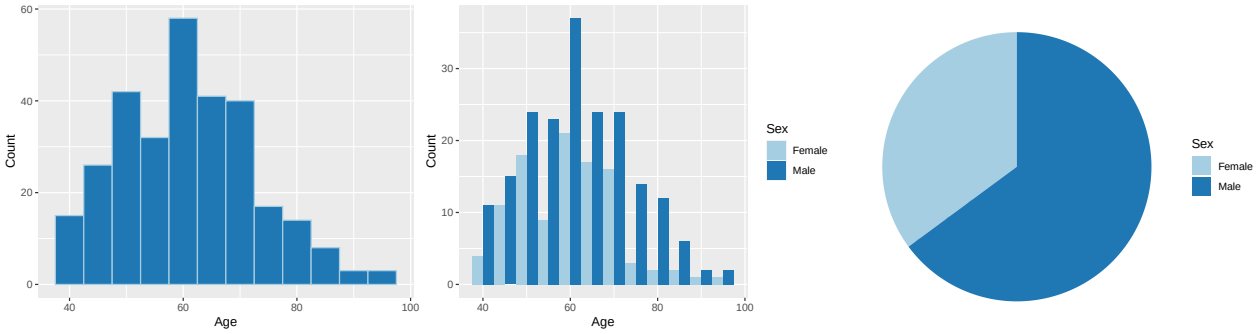

Figure 1: Age and sex distributions

Then we calculated the correlation coefficients between each explanatory variable, and visualized the variance-covariance matrix as a heatmap. We notice that the highest magnitude correlation between variables was 0.44 between sex and smoking. This correlation indicates that there is some correlation between sex and smoking, although this value is not very high so we shouldn't be worried about multi-collinearity in our models. Multicollinearity happens when variables are highly correlated, creating an inflation in the coeffecients and may lead to redundant terms in our model as one variable can predict the other.

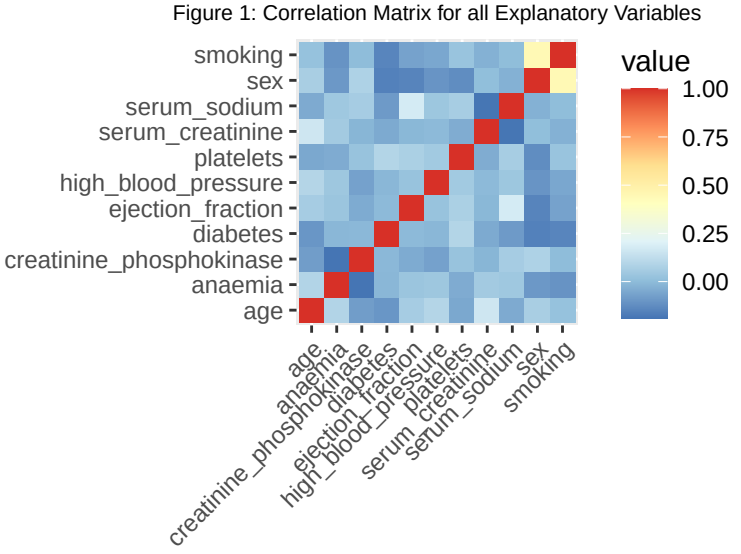

Figure 2: Correlation Matrix for all Explanatory Variables

Model Selection

Based on our choice of death event as our response variable, we decided to use logistic regression for prediction, first we loaded the data and took a cursory look at it, we had 12 explanatory variables and the last column, death event as our response variable. In the model selection step, we used the AIC and BIC (Akaike and Bayesian Information Criterion, respectively) methods, and chose the better selection method between these two.

For testing purposes, first the data was split into two different datasets: a training dataset and a testing dataset. These two initial splits were used to verify all models used so we could compare them fairly against each other.

Method 1: Selection through the Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a criterion that is based on the negative log-likelihood and the quantity of variables used. The AIC penalizes the model as the number of explanatory variables increases. As such, a model with less variables will be less penalized than a model with many variables. When comparing models, a smaller AIC indicates a better overall fit.

Using the step() function we used a stepwise algorithm to choose a model by AIC. The output of this function told us that we should use a 7 variable model with an AIC of 250.87, that uses age, anaemia, creatinine phosphokinase, ejection fraction, high blood pressure, serum creatinine and serum sodium. However, 4 of those variables had a p-value above 0.05, hence these coefficients variables are not statistically significant and we may be able remove them.

We then ran a test to see if we should remove them based on accuracy of the model using our testing data, and comparing this model to subsequent models. By classifying any predicted probabilities over 0.5 as a prediction of death after heart failure, we achieved an accuracy of 80.8%, with the following confusion matrix:

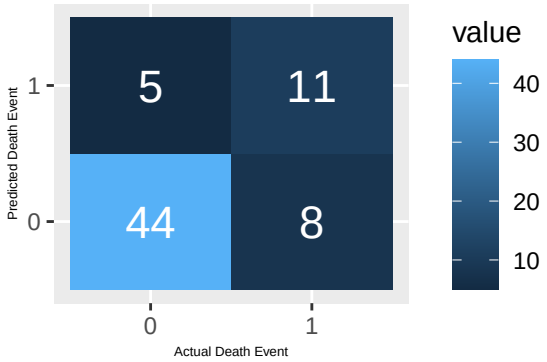

Figure 3: Confusion matrix for Model 1 (AIC)

This confusion matrix told us that the proportion of false positives given that there was a death event was 0.102, and the proportion of false negatives given that there was a death event was 0.421. This was quite high, so it would be good to explore the alternative method of variable selection, which in this case was using the BIC model selection method.

Method 2: Selection through the Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is a criterion that is very similar to AIC as it is based on the likelihood but it places models with many variables more than AIC does. When using the BIC as model selection criteria, we may calculate the posterior probability of each variable and of all models, and select the model with the highest posterior probability.

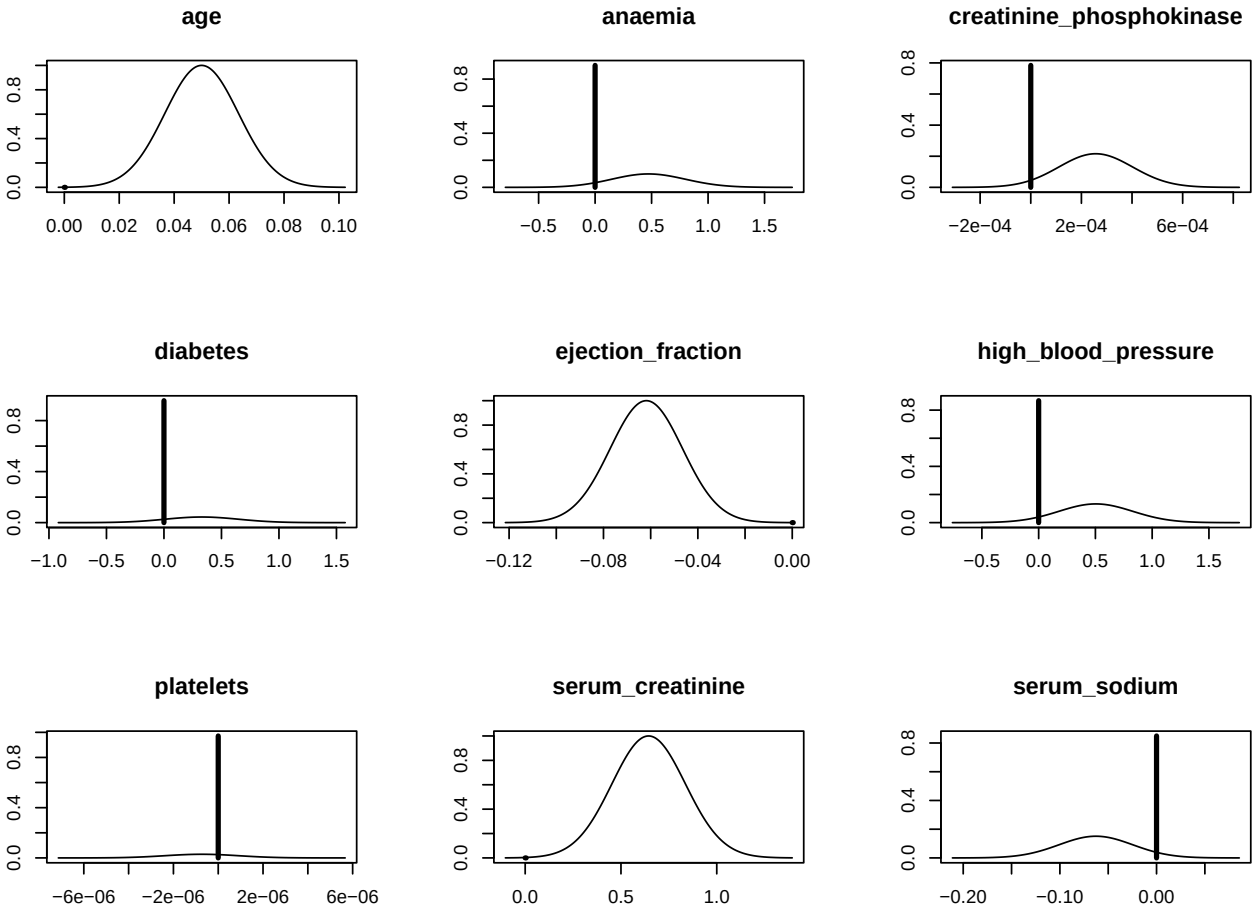

Figure 4: Posterior Distributions for some Explanatory Variables

The posterior distributions displayed above showed that age, ejection fraction and serum creatinine had very small mass at 0, which indicated that these variables should be included in the model, whereas the other variables had a larger mass at 0, indicating they might not be included in the model. We can then looked at the posterior probabilities for our models by showing the value of 'post prob' of 5 models in a histogram.

Based on Bayesian posterior probabilities of the models, the selected model is the model containaing age, ejection fraction and serum creatinine. We can see that this model's posterior probability is about 3 to 5 times higher than the other four models, hence we choose this model and compare it with the previous model chosen by AIC.

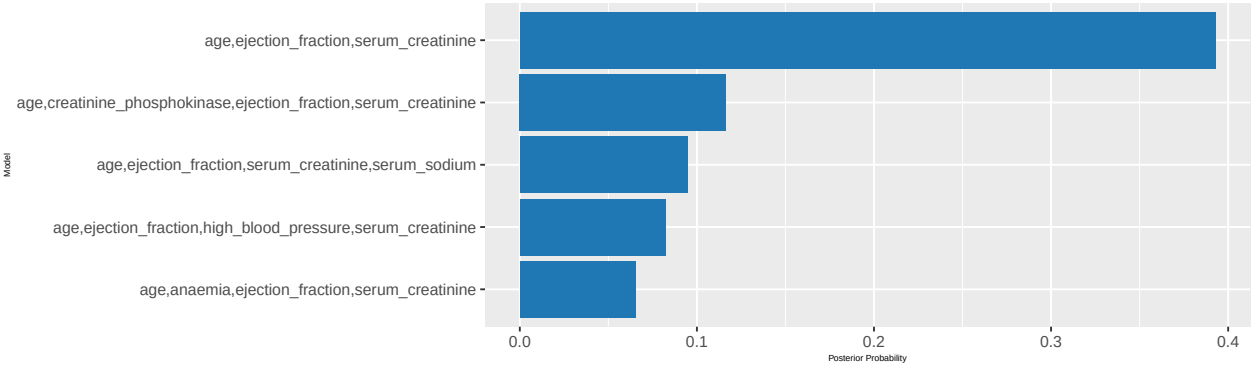

Figure 5: 5 Highest Posterior Probability Models

The formula of the model is given by

$$\hat{\pi} = \frac{e^{-2.349+0.049x-0.062y+0.644z}}{1+e^{-2.349+0.049x-0.062y+0.644z}}$$

where the variables x, y and z represent age, ejection fraction and serum creatinine respectively and $\hat{\pi}$ is the estimated probability of death. We tested the accuracy using the same data sets as we did with the AIC model and we achieved an overall accuracy of 83.8 percent with the following confusion matrix:

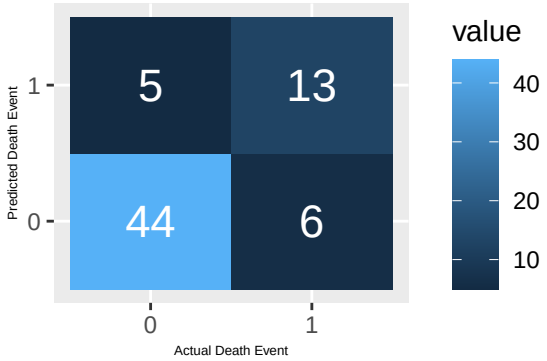

Figure 6: Confusion Matrix for Model 2 (BIC)

Here we have the same proportion of false postives, but the proportion of false negatives is 11 percentage points lower, which is a great improvement from the model found using AIC. Although our second model had less variables included, it actually had a better prediction accuracy than that found using AIC, as well as a lower proportion of false negatives. Hence we should take this one over the model found in Method 1.

Improving the Model through Principal Component Analysis

Principle component analysis allows us to reduce the dimensions of the data by rotating the axes so that the first principle component is now the axis with the most variation in the data, and the second principle component will be the second axis will be an axis that is orthogonal to the first axis.

Doing this allows us to reduce noise in the data and improve prediction rates. When computing the probability of death using this model, we must first apply the transformation that was applied to the training data set to our testing data set so that we guarantee that we are using the same components.

Using the first two components and discarding the third would keep 70% of the variability of the data. Using these two components as explanatory variables for our model leds to a prediction rate of 86.7% and the following confusion matrix:

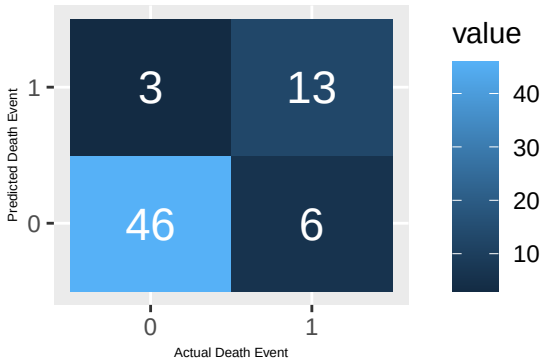

Figure 7: Confusion Matrix for Model 2 after PCA

Using principal component analysis, we were are able to increase the number of true negatives and decrease the number of false positives by 2, which is an increase by 4 percentage points. Overall this was a great improvement, and more reliable and accurate model than that found with the first method.

Conclusion

To reiterate, our research question was: Which variables create the most efficient model to predict a death event in patients who have experienced heart failure? Our project aimed to iteratively build multivariate logistic models with the "heart failure clinical records" dataset in order to arrive at the most efficient relationship to predict a death event following heart failure. With the death event as our response variable and with 12 explanatory variables, we used the AIC and BIC (Akaike and Bayesian Information Criterion, respectively) model selection methods to choose the best model to predict a death event.

The output of the AIC model selection method provides a 7-variable model that uses age, anemia, creatinine phosphokinase, ejection fraction, high blood pressure, serum creatinine, and serum sodium, however, 4 of these variables (age, anaemia, high blood pressure, and serum sodium) had p-values above 0.05 indicating that they were not statistically significant and could be removed from the final model. After classifying any predicted probabilities over 0.5 as a prediction of death after heart failure, this model was able to achieve a prediction accuracy of 80.8% against the testing dataset. In contrast, the output of the BIC model selection method was able to select a model with a prediction accuracy of 83.8% against the testing dataset, selecting a model containing age, ejection fraction and serum creatinine.

When looking at the confusion matrices of the model selected by AIC and BIC respectively, both models shared the same proportion of false positives but the proportion of false negatives for the BIC model was 11 percentage points lower that the AIC model, which was a significant improvement. Furthermore, with the BIC model having only 3 explanatory variables, as opposed to the AIC model's 7 explanatory variables, it had a better prediction accuracy than the AIC model. Coupled with the lower proportion of false negatives, we were able to conclude that the BIC model should be selected over the AIC model. Furthermore, we were able to improve the selected model through principal component analysis, thus improving the model's prediction accuracy to 86.7% and with a confusion matrix indicating an increase in the number of true negatives and a decrease in false positives. In short, this improved model serves as a more reliable and accurate model than that found with the AIC model selection method.

Coming back to our research question of "Which variables create the most efficient model to predict a death event in patients who have experienced heart failure?", we were able to identify the variables that would create the most efficient model to predict a death event in patients who have experienced heart failure, those variables being the age of the patient (in years), ejection fraction (% of blood leaving the heart at each contraction), and serum creatinine (in mg/dL).

Appendix

Akaike Information Criterion

The AIC for a given model with p parameters is defined by

$$AIC = -2(l(\hat{\pi}; y) - p)$$

where $l(\hat{\pi}; y)$ is the loglikelihood of data y evaluated at the parameter values $\hat{\pi}$.

Bayesian Information Criterion

The BIC for a given model with p parameters and n data points is defined by

$$BIC = -2l(\hat{\pi}; y) + p\log(n)$$

where $l(\hat{\pi}; y)$ is the loglikelihood of data y evaluated at the parameter values $\hat{\pi}$.