Tuyen Huynh

Title: Survival of Heart Failure

Abstract:

Although the overall F test shows that the complete model is useful, the probability of mortality of a patient with heart failure is mostly determined by the patient's age (years) and their level of serum sodium present in the blood (mEq/L) when partial F test was used. The other two independent variable, the amount of platelets in the blood and the level of serum creatine, were not as useful individually through the partial F test.

Introduction:

Heart diseases are the world's leading cause of death. The three key factors in heart diseases are high blood pressure, high blood cholesterol, and smoking. There are more factors that play a part in heart failure such as age. For example, if a person is much older, there is a higher chance of them having heart failure issues. Some people might not even realize that they have any heart issues until they experience a sign or a symptom. It would be of utmost importance to understand what truly influences the probability of having a heart failure and the survival rate for it; therefore, using hypothesis tests will be used to see if there is any correlation between the survival rate and the independent variables including one's age, the number of platelets in their blood, the levels of serum creatine and serum sodium found in their blood.

Description of Data:

There are four independent variables: their age, platelets, serum creatinine, and serum sodium. The patient's age will be a factor and measured by years. The platelets in the blood will be measured by kiloplatelets per mL. The levels of serum creatine will be measured by mg per dL and the levels of serum sodium will be measured in mEq per L. The dependent variable would be whether or not the patient died during the follow-up period with 0 being that they survived while 1 being that they died. The data is a clinical record and was gathered through other's research articles.

Description of Statistical Methods:

All tests will be done at a 95% confidence level and R studio was used to calculate all the information/math. To start off the research, the regression model was created, and the normality assumption was assessed by examining the skewness (the degree of symmetry in a distribution) and kurtosis (the heaviness of tails relative to the middle of the distribution) of the data. Then, Cook's distance, leverage statistics and jackknife residuals were used to examine outlier diagnostics and identifying any potential outliers. Afterwards, the overall F test is used to indicate if the complete model with all the independent variables together is useful in predicting the survival of the patients. To go into more depths, the partial F test were performed for all the independent variables separately to see if they are useful in predicting the survival rate on their own. All R codes used will be listed in another document.

Results and Conclusion:

The regression model for the complete model is

$$mortality \;=\; 1.729 + 8.219e^{-3}(age) \;+\; -9.496e^{-8}(platelets) + 1.056e^{-1}(creatinine) \;+ -\,1.486e^{-2}(sodium)$$

. The skewness of the complete model is 0.689; this shows that there are more values above the mean and there is a weak violation of normality assumption. The kurtosis is 2.135, which means that the tail is moderately heavy; it is indicating that there is a violation of normality assumption, but it is difficult to tell and most likely not every severe. Through Cook's distance, leverage statistics and Jackknife residuals, there isn't any noticeable outliers in the data results given since the numbers are quite small. The overall F test indicates that the complete model is useful in predicting the survival of a heart failure patient because the p-value given is less than the alpha which is 0.05; therefore, rejecting the null hypothesis and concluding that there is sufficient evidence to indicate that the complete model is useful in predicting the patient's survival rate. Then onto the partial tests for the usefulness of the independent variables individually. The partial F test for the model with just age concluded that since p-value is less than 0.05, the null hypothesis is rejected and concluding that there is enough evidence to say that age is useful in predicting the survival of the patient. While the partial F test for the model with the number of platelets in the blood shows that the p-value is much greater than 0.05, the null hypothesis is not rejected and there is no sufficient evidence to state that the number of platelets in a patient's blood is useful to predict their survival. That is also the same for the model with the level of serum creatinine in the blood, concluding that there is not enough evidence to state that the level of creatinine is any use in predicting survival rates. The last partial F test is for the model with the level of serum sodium in the blood, the p-value for this one is less than 0.05; thus, rejecting the null hypothesis and since there is enough evidence, concluding that the level of sodium found in the blood is useful in predicting a heart failure patient's survival. Overall, only two out of four independent variables are found useful in predicting mortality caused by heart failure on their own; thus, showing that age and the level of sodium in the blood are the most important factors to look at when trying to predict the probability of survival for a patient with heart failure.

Citation

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). (link)