Bayesian Modeling and Factor Analysis for Heart Failure Risk

STATS 451 Winter 2024 Group Project

Jiayi Ji, Samuel Tan, Ziming Wang

April 27, 2024

Outline:

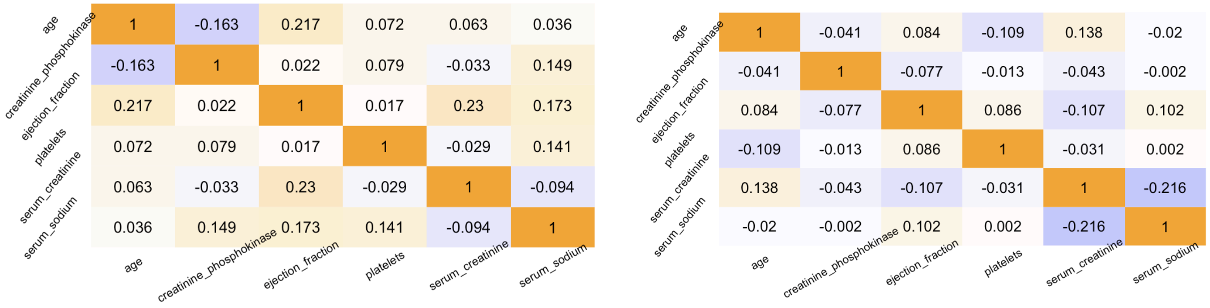
1: Introduction

Cardiovascular diseases (CVDs) stand as the predominant cause of death around the globe, contributing to approximately 31% of all mortalities. Among the numerous conditions CVDs encompass, heart failure is a critical symptom indicative of the underlying disease severity. The data of interest is sampled from medical records of 299 heart failure patients at the Faisalabad Institute of Cardiology and Allied Hospital in Pakistan, collected between April and December 2015. In the dataset, the response is DEATH_EVENT, a binary variable taking a value of 1 if the patient died, and 0 if not. Predictors include participants' demographics, health conditions, and lifestyle factors. In detail, the utilized demographic predictors are sex and age. Health conditions include anemia, creatinine phosphokinase, diabetes, ejection fraction, hypertension or not, platelets in the blood, serum creatinine level, and serum sodium level. Lifestyle predictor is whether patients smoke or not. Also we have a time predictor that represents the follow-up period. In total, we have 203 patients still alive and 96 patients died. We aim to 1) apply Bayesian methods to measure the influence of various determinants on the likelihood of heart failure across diverse individuals and 2) pinpoint predictors that markedly elevate heart failure risk to inform early detection strategies and facilitate the screening of at-risk populations.

2: Exploratory Data Analysis

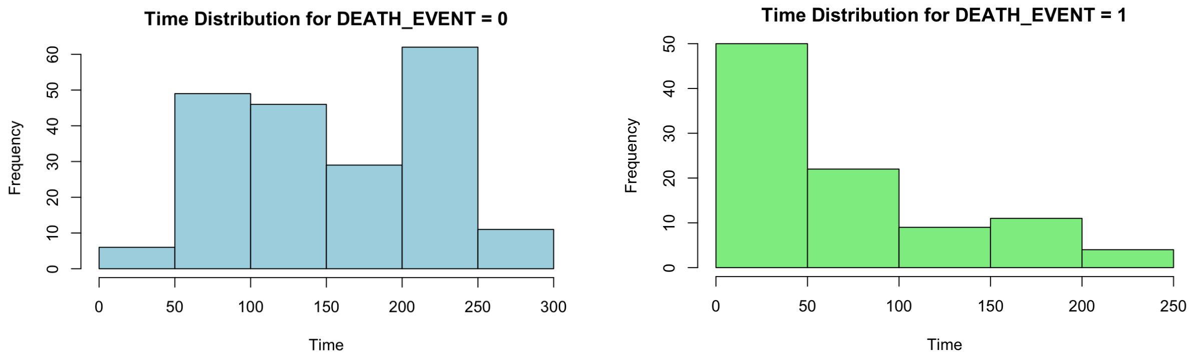
The dataset is a simple random sample from the population of heart failure patients during that period and region. It's well-curated without obvious defects but with limitations lie in its regional focus which may not fully represent the broader heart failure patient population due to local healthcare system variables or environmental factors. There are no missing values since they come from well-documented clinical records of hospitals and research centers.

We utilize Exploratory Data Analysis (EDA) to explore and visualize our dataset to understand relationships between death events and other predictors, and also detect any potential collinearity between predictor variables.



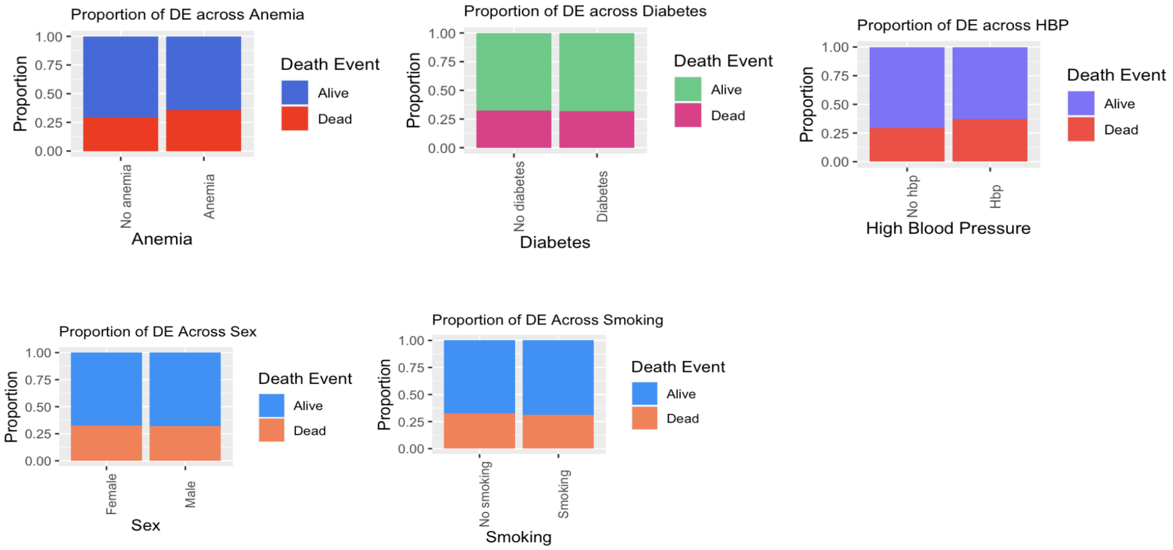
(Figure 1a (left: death event Dead) & Figure 1b (right: death event Alive))

Correlation coefficients showed different trends on different statuses of death events. There are more positive coefficients on dead status while more negative coefficients on alive status. The most negative coefficients are -0.163 and -0.216, which are between age and creatinine phosphokinase and serum creatinine and serum sodium. It might indicate that higher age tends to have lower creatinine phosphokinase and also the negative relationship between serum creatinine and serum sodium. These two heatmaps do not have coefficients really close to 1 or -1, there is no statistically significant collinearity among our selected variables.



(Figure 2a (left: Time distribution for alive) & Figure 2b (right: Time distribution for dead))

The time variable represents the follow-up period researchers followed the patients. Based on histograms from Figure 2a and Figure 2b, there's obvious trends that time correlates with death event response. Naturally, researchers follow the surviving patients for a longer time. Therefore, we decided to remove the time variable from our dataset.



((Figure 3a (Anemia), Figure 3b (Diabetes), Figure 3c (HBP), Figure 3d (Sex), Figure 3e (Smoking)))

In Figure 3a, participants who have anemia have a higher proportion of dead. Also, the proportion of dead is higher among participants who have high blood pressure. From Figure 3b, Figure 3d and Figure 3e, patients' diabetes status, sex, and smoking don't present obvious differences in the proportion of dead.

3: Methods

We will be implementing a Bayesian Logistic Regression in R, using the RStan package. MCMC sampling algorithm will be used to obtain the posterior samples of each beta coefficient corresponding to each feature. The 95% credible interval of beta coefficients will then be obtained from the sample, this represents that our unknown parameter, or in this case, the beta coefficients, have a 95% probability that they will lie between this interval. If this interval does not contain zero, it is highly likely that our unknown beta coefficients are not zero, and thus significant in affecting the risk of heart failure. In the following subsections, We will examine the prior, likelihood, and posterior we have chosen for this model and explain the rationale behind these choices.

3.1 Likelihood (Proposed Model)

The response variable in our dataset, DEATH_EVENT, is binary, with 1 representing death, and 0 representing survival. We want to use 11 features to predict DEATH_EVENT, and interpret the influence of each feature on DEATH_EVENT. The nature of this problem is a binary classification task and the emphasis on interpretation of our model makes Logistic Regression a very reasonable choice. The likelihood model is the following:

$$y[i]|\alpha, \beta, x[i] \sim \text{Bernoulli}(\text{logit}^{-1}(\alpha + x[i]^T\beta))$$

Figure 4: Likelihood (Proposed Model)

Where subscript i represents different trials, alpha represents the interception of the linear predictor function, and beta is a vector of regression coefficients. The inverse logit transformation ensures that the output probabilities lie between 0 and 1. Bernoulli distribution then uses these probabilities to determine the likelihood of a death event occurring, which directly ties into our predictive model.

3.2 Prior

We don’t have any prior knowledge of the regression coefficients, so we will be using a non-informative normal prior for all the coefficients. We will center the normal prior with mean 0, implying that prior to seeing the data, it is equally likely for the influence on a predictor to be positive or negative. We chose the variance to be 10, which is relatively large considering the scale of the data. This represents our weak belief about the parameters in our Bayesian model.

Formally, our priors are:

$$\begin{aligned} \alpha &\sim \text{Normal}(0, 10) \\ \beta_j &\sim \text{Normal}(0, 10) \quad \text{for each } j = 1, \ldots, 11 \end{aligned}$$

Figure 5: Prior distributions

Where alpha represents the interception of the linear predictor function and beta_j represents the regression coefficient for each feature used to predict DEATH_EVENT. After the removal of the “time” variable, we are left with 11 features, so “j” ranges from 1 to 11, representing each feature respectively.

3.3 Posterior

After defining the likelihood and choosing our priors, we can then use the Bayes formula to calculate the posterior distribution. The posterior distribution is proportional to the likelihood times the prior. Considering our likelihood and prior, we can write the posterior distribution as follows:

$$p(\alpha, \beta | y, x) \propto p(\alpha) \prod_{j=1}^{K} p(\beta_j) \times \prod_{i=1}^{N} p(y[i] | \alpha, \beta, x[i])$$

Figure 6: Posterior Distribution

The RStan implementation of this Bernoulli-Logit model can be found in the R code attached with the submission.

4: Results

Figure 9 is a summary of the Bayesian Logistic Regression Model’s coefficient fitting using RStan, displaying the central tendency(mean) and uncertainty(variance) of the coefficients' posterior distribution. The percentiles represent the Bayesian credible intervals for the coefficients, providing a range where the parameter values are most likely to lie. Several coefficients show a statistically significant impact on the heart failure prediction indicated by their credible interval. For example, the coefficient of age has a 95% credible interval between 0.049 and 0.125, with a mean equal to 0.09. Its credible interval does not include 0, suggesting that there is strong evidence that the coefficient of age is significantly different from zero under the model's assumptions. Another regression coefficient of interest is the coefficient of the variable ef (ejection fraction). The 95% credible interval of ejection fraction ranges from -0.13 to -0.06, with a mean equal to -0.09. Again, its credible interval does not include 0, suggesting that there is strong evidence that the coefficient of age is significantly different from zero under the model's assumptions. Figure 10 shows the posterior distribution of some selected predictor coefficients including age, and most of the posterior samples do not cross zero. Creatinine Phosphokinase(CPK), on the other hand, displays a tight credible interval around zero, indicating that its impact on heart failure prediction is likely not statistically significant.

4.1 Convergence Analysis

In the previous section, we identified beta_age and beta_ef to be two regression coefficients of interest. To check for the convergence of our MCMC algorithm, we will examine the Rhat of these two coefficients. Below is the table displaying Rhat values for beta_age and beta_ef, as well as the traceplot.

	beta_age	**beta_ef**
Rhat	**1.03**	**1.07**

Figure 7: Rhat of beta_age and beta_ef

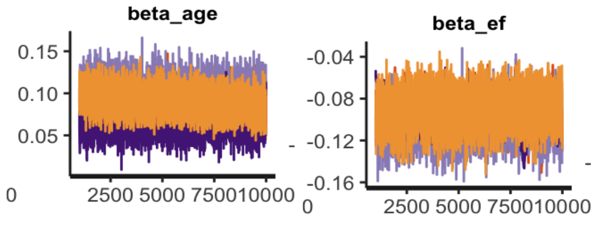
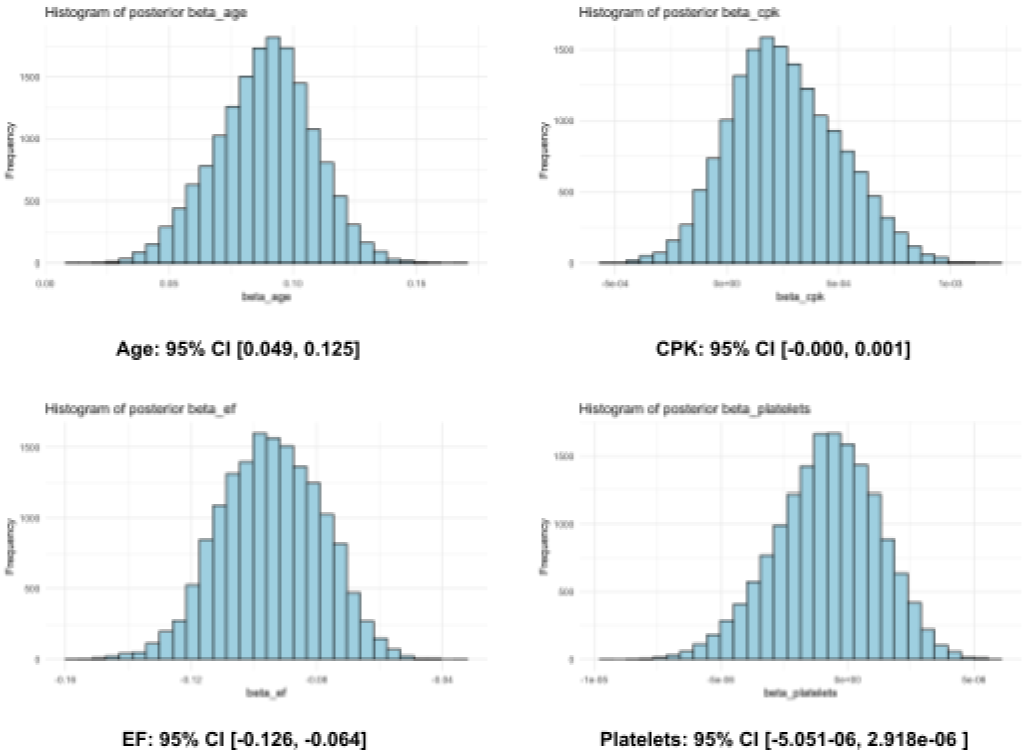


Figure 8a(left), 8b(right): Traceplot for beta_age (left) and beta_ef (right)

The Rhat of beta_age is 1.03, which is close to 1 and under the threshold of 1.05, indicating the convergence of our MCMC algorithm. The Rhat of beta_ef is 1.07, which is slightly above the threshold of 1.05. This might suggest a potential lack of convergence or the need for further iterations.

Parameter	**Mean**	**SD**	**2.5%**	**25%**	**50%**	**75%**	**97.5%**
alpha	**-0.11**	**1.16**	**-1.49**	**-1.15**	**-0.06**	**1.07**	**1.11**
age	**0.09**	**0.02**	**0.05**	**0.08**	**0.09**	**0.10**	**0.13**
anaemia	**-0.17**	**1.24**	**-1.64**	**-1.14**	**-0.34**	**0.60**	**1.65**
cpk	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
diabetes	**0.10**	**0.66**	**-0.66**	**-0.40**	**-0.06**	**0.49**	**1.13**
ef	**-0.09**	**0.02**	**-0.13**	**-0.11**	**-0.09**	**-0.08**	**-0.06**
hbp	**-0.01**	**1.34**	**-1.46**	**-1.24**	**-0.20**	**1.07**	**1.80**
platelets	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
creatinine	**-0.21**	**1.37**	**-1.64**	**-1.19**	**-0.58**	**0.18**	**2.06**
sodium	**-0.01**	**0.01**	**-0.03**	**-0.02**	**-0.01**	**-0.01**	**0.00**
sex	**-0.63**	**0.96**	**-1.82**	**-1.12**	**-0.84**	**-0.27**	**0.90**
smoking	**-0.63**	**0.39**	**-1.31**	**-0.73**	**-0.47**	**-0.37**	**-0.26**

Figure 9: Summary of posterior distribution of beta coefficients



Age: 95% CI [0.049, 0.125]

CPK: 95% CI [-0.000, 0.001]

EF: 95% CI [-0.126, -0.064]

Platelets: 95% CI [-5.051-06, 2.918e-06]

Figure 10: Selected Posterior distributions of some beta coefficients

5: Conclusion

By sampling from the posterior distribution, we have discovered that the 95% credible interval of regression coefficients for the variable age and ef (ejection fraction) does not include 0. This suggests that there is strong evidence that the coefficient of age and ef are significantly different from zero under the model's assumptions. We want to note that ejection fraction is a measurement of the percentage of blood leaving the heart each time it squeezes. A low ejection fraction suggests that the patient's heart isn't pumping enough blood and may be failing, so it is reasonable to see that the regression coefficient for this variable is significant and negatively correlated with heart failure risk.

For age variables, we can see that the regression coefficient is significant and positively correlated with heart failure risk. This is also a reasonable conclusion as humans' hearts tend to deteriorate as their age increases.

Overall, we have concluded that age and ejection fraction are the most significant indicators of heart failure according to our analysis. The findings have the potential to enhance the prediction of heart failure risk and to more effectively identify populations that are at an increased risk of developing heart failure.

6: Data Source and Reference

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).

(https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-)

7: Individual Contribution

Jiayi Ji: Exploratory Data Analysis, Data Cleaning, and Visualization, wrote parts 1 and 2 (Introduction, Data) of the final report.

Samuel Tan: Model Design and Results Interpretation. Wrote parts 3 and 5 (Method, Conclusion) of the final report.

Ziming Wang: Model Implementation, Tuning, and Visualization. Wrote part 4 (Result) of the final report.