BIOSTAT 682 Final Project

Bayesian Inference on Heart Failure Data

Mukai Wang 98830336

1 Introduction

1.1 Background

Cardiovascular diseases(CVD) are the number one cause of death globally. According to WHO, heart failure is responsible for 85% of all the CVD deaths[1]. Heart failure cases are also much more prevalent in low-income and middle-income countries[1]. Finding potential indicators for heart failure is critical for saving lives. This project is a case study of heart failure patients in Pakistan and aims at extracting key indicators for death events and time-to-death.

1.2 Data Summary

The dataset is an electronic health record of heart failure patients in Pakistan collected in 2015[2]. There were 299 patients. All the patients were more than 40 years old, having left ventricular systolic dysfunction and falling in NYHA class III and IV[2]. The followup time ranged from 4 to 285 days. 96 deaths were observed. There are 13 columns in the dataset:

- `DEATH EVENT`: A binary variable indicator of death. This is the primary variable of interest.
- `time`: Followup time for each patient. This is another variable of interest.
- `Age`: The patients' ages range from 40 to 95 years old.

- `Anemia`: A binary variable indicating if a patient has anemia.
- `Creatine Phosphokinase(CPK)`: Creatine kinase measured in mcg/L.
- `Diabetes`: A binary variable indicating if a patient has diabetes.
- `Ejection Fraction (EF)`: Measurement of blood leaving the heart at each contraction (in percentage).
- `High Blood Pressure`: A binary indicator of high blood pressure.
- `Platelets`: Number of platelets per 1ml blood.
- `Serum Creatinine`: Creatinine concentration in blood measured in mg/dL.
- `Serum Sodium`: Sodium level in blood measured in mEq/L.
- `Sex`: Binary variable, 1 for men and 0 for women.
- `smoking`: A binary variable indicating if a patient smokes.

1.3 Project Goal

The project goal is to identify factors that have significant association with heart failure. I will perform logistic regression to predict whether a patient would die from heart failure. I will also perform parametric survival analysis on time-to-death based on weibull regression. I will do variable selection when fitting models in both analysis. Both analysis will be done with Bayesian inference.

2 Methods

2.1 Data Preprocessing

Because values in different columns have very different ranges and magnitudes, I need to preprocess the data first. Variables `age`, `ejection fraction`, `platelets` and `serum sodium` are centered. `Creatine Phosphokinase` and `Serum Creatinine` are taken log of 10.

To compare the performance of inference methods, I split the dataset into a training set and a test set. The training set contains 240 individuals and the test set contains 59 individuals. One third of both the training set and test set have observed death events. The model fitting will all be done with the training set and the model performance will be evaluated based on the test set.

2.2 Death Event Prediction

I will use logistic regression to predict whether a patient dies from heart failure. The model is[3] (I use $\boldsymbol{\Delta}$ as the binary vector of death events for all the patients)

$$\log\left(\frac{\pi(\Delta_i)}{1-\pi(\Delta_i)}\right) = X_i^\top \boldsymbol{\beta} \qquad i = 1, 2, \cdots, n \tag{1}$$

n stands for the total number of patients. Vector X_i contains the intercept. The likelihood can be written as

$$\pi(\boldsymbol{\Delta}; \boldsymbol{X}, \beta) = \prod_{i=1}^{n} \left\{ \left(\frac{\exp(X_i^\top \beta)}{1+\exp(X_i^\top \beta)}\right)^{\Delta_i} \left(\frac{1}{1+\exp(X_i^\top \beta)}\right)^{1-\Delta_i} \right\} \tag{2}$$

In Bayesian analysis setting, I need to set a prior for coefficient vector $\boldsymbol{\beta}$. A common choice is multivariate normal distribution $\boldsymbol{\beta} \sim N(\boldsymbol{b}, \Sigma)$[4]. Then we can have the posterior distribution

$$\pi(\beta | \boldsymbol{\Delta}, \boldsymbol{X}) \propto \pi(\boldsymbol{\Delta}; \boldsymbol{X}, \beta) \cdot \exp\left\{-0.5(\boldsymbol{\beta} - \boldsymbol{b})^\top \Sigma^{-1} (\boldsymbol{\beta} - \boldsymbol{b})\right\} \tag{3}$$

The dimension of $\boldsymbol{\beta}$ has to be decided through variable selection. I am going to try two ways:

1. The first way is best subset selection. In this approach, I will exhaustively try all the combinations of 11 covariates in the dataset. I will select the best model out of the 2047 candidates based on the lowest value of DIC[4].

   For a model with p covariates, I will set the prior distribution parameter of $\boldsymbol{\beta}$ to satisfy

   $$\beta_j \sim N(0, 100) \qquad j = 1, 2, \cdots p$$

   This essentially means that **b** is a vector of zeros and Σ is a diagonal matrix with value 100 for all the diagonal terms. The choice is based on my observation of coefficient values after I fit a logistic regression with all the 11 covariates using the frequentist way. Since all the coefficients(including intercept) have an absolute value smaller than 10, I believe assigning a normal distribution with mean of 0 and standard deviation of 10 to each covariate effect is a reasonable prior.

2. The second way is Bayesian lasso regression[5]. Recall that in frequentist settings, given a well defined linear predictor $\boldsymbol{\eta}$ and a link function for generalized linear model

$$\boldsymbol{\eta} = \mu\mathbf{1} + \boldsymbol{X}\beta \qquad E[\boldsymbol{y}|\boldsymbol{X}] = g^{-1}(\mu\mathbf{1} + \boldsymbol{X}\beta)$$

The Lasso estimation of β will be[6]

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\beta}[\log \pi(\boldsymbol{y}|\mu, \beta) - \lambda\|\beta\|_1] \tag{4}$$

In the Bayesian setting, the ℓ_1 penalty term can be understood as a mixture of hierarchical prior distribution[7]

$$\beta_j \sim N(0, \sigma_j^2) \qquad \sigma_j^2 \sim \exp(\lambda) \qquad j = 1, \cdots, p \tag{5}$$

p stands for the total number of covariates. In our logistic regression setting, we can get the best estimator $\hat{\beta}$ and $\hat{\Sigma}$ by maximizing their marginal posterior distribution[7]

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \text{argmax}_{\beta} \log \pi(\beta|\boldsymbol{\Delta}, \boldsymbol{X}, \mu, \lambda) \\ &\approx \text{argmax}_{\beta} \log \int \left[\pi(\boldsymbol{\Delta}|\mu, \beta) \cdot (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}\beta^{\top}\Sigma^{-1}\beta\right) \cdot \prod_{j=1}^{p} \lambda \exp\left\{-\lambda\sigma_j^2\right\}\right] d\Sigma\end{aligned} \tag{6}$$

$$\begin{aligned}\hat{\Sigma} &= \text{argmax}_{\Sigma} \log \int \left[\pi(\boldsymbol{\Delta}|\mu, \beta) \cdot (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}\beta^{\top}\Sigma^{-1}\beta\right) \cdot \prod_{j=1}^{p} \lambda \exp\left\{-\lambda\sigma_j^2\right\}\right] d\beta \\ &= \text{argmax}_{\Sigma} \left[\log \pi(\Delta|\mu, \Sigma) - \sum_{j=1}^{p} \lambda\sigma_j^2 + c\right] \qquad \text{(c is a constant)}\end{aligned} \tag{7}$$

Because all the σ_j^2 are non-negative, the marginal likelihood maximization process might shrink some of the variances to zero. This means the corresponding β_j must be zero, fulfilling the lasso shrinkage functionality. The value of λ needs to be prespecified, and it can be decided based on hyperparameter tuning and cross validation.

2.3 Time-to-Death Prediction

In this part I will take into consideration the time-to-death information. I choose Weibull Accelerated Failure Time(AFT) model to model the survival time for each patient. The model can be expressed either in failure time or survival probability[8],

$$\log(t_i) = X_i^\top \boldsymbol{\beta} + \sigma\epsilon_i \qquad S_T(t_i) = \exp\left[-\left(\frac{t_i}{\exp(X_i^\top \boldsymbol{\beta})}\right)^{1/\sigma}\right] \qquad i = 1, 2, \cdots, n \tag{8}$$

$\boldsymbol{\beta}$ and σ are parameters of interest. ϵ_i follows an extreme value distribution.

There are multiple plausible estimates for survival time. I am going to choose a form called "minimum prediction error survival time"[9]. It's mathematically denoted as

$$\hat{t}_i = \left[\frac{\frac{2}{\sigma}\log(k)}{k^{\frac{1}{\sigma}} - k^{-\frac{1}{\sigma}}}\right]^{\sigma} \exp(X_i^\top \boldsymbol{\beta}) \tag{9}$$

The k is an integer of users' choice. In terms of prediction, $\hat{t}_i/k < t_i < k\hat{t}_i$ is considered correct.

The likelihood of Weibull model can be denoted as[8]

$$\begin{aligned}\pi(\boldsymbol{t}, \boldsymbol{\Delta}; \boldsymbol{\beta}, \sigma) &= \prod_{i=1}^{n} [f_T(t_i)]^{\Delta_i} [S_T(t_i)]^{1-\Delta_i} \\ &= \prod_{i=1}^{n} \left\{\frac{1}{\sigma}\left(\frac{t_i}{\exp(X_i^\top \boldsymbol{\beta})}\right)^{1/\sigma} \exp\left[-\left(\frac{t_i}{\exp(X_i^\top \boldsymbol{\beta})}\right)^{1/\sigma}\right]\right\}^{\Delta_i} \left\{\exp\left[-\left(\frac{t_i}{\exp(X_i^\top \boldsymbol{\beta})}\right)^{1/\sigma}\right]\right\}^{1-\Delta_i}\end{aligned} \tag{10}$$

In Bayesian setting, we need to specify a prior distribution for $\boldsymbol{\beta}$ and σ. A plausible choice for $\boldsymbol{\beta}$ is a multivariate normal distribution $\boldsymbol{\beta} \sim N(\boldsymbol{b}, \Sigma)$ (same as that in logistic regression) and a plausible choice for σ is to choose an uninformative uniform prior for $1/\sigma$, $\pi(1/\sigma) \propto 1$. The form of the posterior distribution for $\boldsymbol{\beta}$ and σ will look like

$$\pi(\boldsymbol{\beta}, \sigma | \boldsymbol{t}, \boldsymbol{\Delta}) \propto \pi(\boldsymbol{t}, \boldsymbol{\Delta}; \boldsymbol{\beta}, \sigma) \cdot \exp\left\{-0.5(\boldsymbol{\beta} - \boldsymbol{b})^{\top}\Sigma^{-1}(\boldsymbol{\beta} - \boldsymbol{b})\right\} \tag{11}$$

I will carry out the best subset selection and pick the model with the lowest DIC out of 2047 candidates. To have a sense of the range of parameter values, I fit the weibull model with all the 11 covariates using the frequentist method. Because all the β_i have an absolute value smaller than 10, I decide to set $\boldsymbol{b}$ to be a vector of zeros and Σ to be a diagonal matrix with all the diagonal terms being 100. This is the same setting as that for logistic regression. As to σ, since I observe a value of about 1 from the full model using frequentist way, I decide to assign a Uniform(0,5) to $1/\sigma$ as the prior distribution.

3 Results

3.1 Death Event Prediction

The Bayesian estimation from the best subset selection and lasso are presented in table 1 together with the best subset selection and lasso output under frequentist scheme.

Best Subset (Frequentist, AIC)		Lasso (Frequentist)		Best Subset (Bayesian, DIC)		Lasso (Bayesian)	
Name	Value (with SD)	Name	Value	Name	Value (with SD)	Name	Value
Intercept	-1.186(0.269)	Intercept	-1.018	Intercept	-1.242(0.254)	Intercept	-1.08
Age	0.038(0.014)	Age	0.024	Age	0.037(0.014)	Age	0.022(0.01)
EF	-5.37(1.46)	EF	-4.01	EF	-5.625(1.319)	EF	-4.00(1.155)
Creatinine	3.92(0.91)	Creatinine	3.168	Creatinine	4.458(0.899)	Creatinine	3.125(0.732)
Anemia	0.629(0.332)	Anemia	0.209	Anemia	0.582(0.323)	Sodium	-0.020(0.020)
CPK	0.635(0.330)	CPK	0.255	CPK	0.628(0.327)		
Sodium	-0.056(0.035)	Sodium	-0.030				
		HBP	0.130				
		Sex	-0.112				

Table 1: Logistic Regression Coefficients for four different inference methods. `EF` stands for "Ejection Fraction". `CPK` Stands for "Creatinine Phosphokinase". `HBP` stands for "High Blood Pressure". Standard errors are not available for the frequentist lasso methods from R output.

We can check out the model performance on the test set based on their predicted probability. A rigorous way to evaluate binary prediction performance is to calculate the area under the Receiver Operating

Characteristics(ROC) curve[10]. The ROC curves are shown in figure 1. All four inference methods have good prediction performance. The Bayesian lasso method achieves the highest score with the fewest covariates.

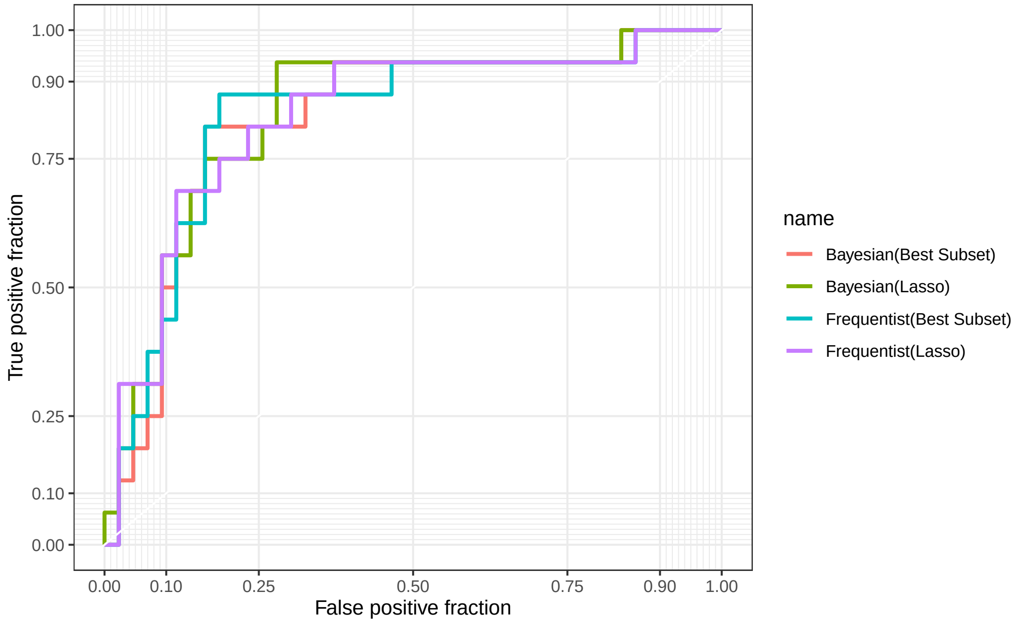

Figure 1: Receiver operating characteristics curve for four different inference Methods. The calculated areas under the curve are: 0.8256 for best subset selection using Bayesian method, 0.8314 for best subset selection using frequentist method, 0.8372 for Bayesian lasso method, 0.8328 for frequentist lasso method.

The significant factors are the age, ejection fraction and serum creatinine concentration. I can interpret the coefficient values for these three factors based on the Bayesian lasso output as:

- Adjusted for other covariates, patient A's death odds is about 25% higher than patient B if patient A is older than patient B by one year.
- Adjusted for other covariates, patient A's death odds is 33% lower than patient B if patient A's ejection fraction is higher than patient B's ejection fraction by an absolute value of 10%.
- Adjusted for other covariates, if a patient's serum creatinine level scales up 10 times, that patient's death odds will scale up by about 22.8 times

3.2 Time-to-death Prediction

As mentioned in the methods section, I carry out best subset selection for a weibull survival model with both the Bayesian and frequentist way. The models are summarized in table 2.

Best Subset (Frequentist, AIC)		Best Subset (Bayesian, DIC)	
Name	Value (with SD)	Name	Value (with SD)
σ	1.06(0.103)	σ	1.14(0.11)
Intercept	6.59(0.27)	Intercept	6.97(0.273)
Age	-0.03(0.01)	Age	-0.033(0.012)
Anemia	-0.484(0.242)	Anemia	-0.337(0.214)
EF	4.178(1.168)	EF	4.151(1.062)
Creatinine	-2.436(0.563)	Creatinine	-3.022(0.666)
HBP	-0.384(0.242)	HBP	-0.378(0.221)
Sodium	0.040(0.024)	Platelets	-1.653(1.289)
CPK	-0.413(0.258)	Diabetes	-0.303(0.257)
		Smoking	0.213(0.321)

Table 2: Weibull model for time-to-death prediction estimated in both Bayesian and frequentist ways

I use formula 9 to decide whether the predicted survival time is accurate given the observed follow up time. Among the 16 individuals that have observed death events in the test set, the frequentist inference method successfully predicts survival time for 6 patients and the Bayesian method successfully predicts survival time for 5 patients. The correct predictions all belong to patients with long survival time(longer than 70 days).

Even though the weibull model contains more covariates than logistic regression, the significant factors are the same. They are the age, ejection fraction and serum creatinine concentration. I can interpret their effects as:

- Adjusted for other covariates, patient A will have an expected survival time shorter than patient B by about 3% if patient A is older than patient B by one year.
- Adjusted for other covariates, patient A will have an expected survival time longer than patient B by about 51% if patient A's ejection fraction is higher than patient B's ejection fraction by an absolute value of 10%.

- Adjusted for other covariates, if a patient's serum creatinine level scales up 10 times, the patient's expected survival time will scale down by about 96%.

4 Discussion

This project looks into death caused by heart failure from both the death probability perspective as well as the survival time perspective. I use logistic regression to model death probability and weibull model to model survival time. Both the logistic regression and the weibull model detect age, ejection fraction and serum creatinine concentration to be significant factors. Older patients have higher death risk. Patients with higher ejection fraction have lower death risk. Patients with high serum creatinine level have very high death risk. These findings agree with previous publications on the same dataset[11].

In terms of model performance, the logistic regression achieves high prediction accuracy for both frequentist and Bayesian methods. The weibull model for survival time is much less accurate regardless of the inference method. There are three explanations. The first is that the sample size is too small. We only have 299 individuals in total, which make survival time prediction challenging. The second is that weibull model might not be the best survival model for this dataset. There may be other parametric or semiparametric models that work for this dataset better. The third is potential overfitting. The weibull model picks up more factors than logistic regression, and a lot of factors are deemed not significant based on their P values.

5 Computation Notes

All the analysis are done in R. The best subset selection in Bayesian way is done through `R2jags` package which is the R interface for `JAGS`. The Bayesian lasso regression is done with `EBglmnet` package. The frequentist best subset selection for logistic regression is done with `glm` function in the basic `stats` package. The frequentist lasso regression is done with `glmnet` package. The best subset selection is computationally expensive and is carried out by submitting job arrays on University of Michigan Biostatistics high performance computing(HPC) cluster. Part of the codes for Bayesian analysis can be found on github.

References

[1] Cardiovascular disease(cvd). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] Tanvir Ahmad, Assia Munir, Sajjad Bhatti, Muhammad Aftab, and Muhammad Ali Raza. Survival analysis of heart failure patients: A case study. *PLOS ONE*, 12:e0181001, 07 2017.

[3] Annette J. Dobson. *An introduction to generalized linear models / Annette J. Dobson*. Chapman Hall/CRC Boca Raton, 2nd ed. edition, 2002.

[4] Tim Johnson. Biostat 682 fall 2020 lecture notes.

[5] Trevor Park and George Casella. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008.

[6] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

[7] Ebglmnet vignette. https://cran.r-project.org/web/packages/EBglmnet/vignettes/EBglmnet_intro.pdf.

[8] Enwu Liu and Karen Lim. Using the weibull accelerated failure time regression model to predict time to health events. *bioRxiv*, 2018.

[9] R. Henderson, M. Jones, and J. Stare. Accuracy of point predictions in survival analysis. *Statistics in medicine*, 20 20:3083–96, 2001.

[10] Tom Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874, June 2006.

[11] Davide Chicco and 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(1):16, Feb 2020.