A case study on survival data of passengers on Titanic

Zhicong Chu

Abstract

The sinking of Titanic is one of the most famous tragedies in the human history. It has been more than a century since the calamity. But it is always valuable to look back, to learn from the history and to avoid repeating of the tragedy. The survival data of passengers on Titanic precisely gives us such an opportunity. In this study, I modeled the survival data through three different methods: Bayesian logistic regression, classical logistic regression and artificial neural network. The performance of the models were evaluated through prediction and validation on a test data set. The results gave some insight on the rescue operation, indicating that certain groups of passengers were more likely to survive the disaster than others such as women, the younger and passengers with higher class ticket and more Siblings or spouse on board.

Introduction

In the early morning of 15 April 1912, RMS Titanic sank into the Atlantic Ocean after colliding with an iceberg. Thousands of people lost their lives in the shipwreck. One of the most important reasons that the calamity led such great loss in lives is the insufficiency of lifeboats. It cannot be denied that some element of luck was involved in surviving the tragedy. But from the perspective of humanitarian, some groups of people might and should have greater chance to be rescued such as women and children. So I hypothesized that woman and children on board were more likely to survive the sinking of Titanic.

The full dataset have 12 variables on 714 observations after I delete the observations with missing values.

Variable Name	Description
Survived	Survived (1) or died (0)
Pclass	Passenger's class
Name	Passenger's name
Sex	Passenger's sex
Age	Passenger's age
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Fare
Cabin	Cabin
Embarked	Port of embarkation

Table 1. A brief description of the variables listed in the full data set.

I select 5 variables that might have effect on the probability of survival. They are listed as follows: The first variable Pclass reveals the three different categories of the cabins that is first, second and third class. Followed by Sex showing the gender of each person. Then it has the variable Age, recording the age of the passengers at the shipwreck. The next variables SibSp gives the information of number of siblings and spouses on board of each passenger. And the last variable Parch reveals the number of parents and children on board of each person. The variable Survived represents the survival status of each person and is also our binary response.

```
head(data)
```

```
##     Pclass  Sex Age SibSp Parch Survived
## 24       1 male  28     0     0        1
## 1        3 male  22     1     0        0
## 223      3 male  51     0     0        0
## 666      2 male  32     2     0        0
## 222      2 male  27     0     0        0
## 386      2 male  18     0     0        0
```

Figure 1. The first several rows of the dataset after variable selection. Survived is the response. Other 5 variables thought to have effect on the survival status were selected and a subset of the full dataset was generated as our dataset of interest using these 5 predictor variables together with the response variable.

Overview of data and data split

1) Descriptive statistics

The data set contains 5 predictor variables with 2 categorical variables and 3 continuous variables. And the predicted variable is binary. All the input variables are approximately on the same scale. So feature scaling and mean normalization is not very necessary.

```
summary(data)
```

```
##  Pclass      Sex           Age            SibSp           Parch      
##  1:186   female:261   Min.   : 0.42   Min.   :0.0000   Min.   :0.0000  
##  2:173   male  :453   1st Qu.:20.12   1st Qu.:0.0000   1st Qu.:0.0000  
##  3:355                Median :28.00   Median :0.0000   Median :0.0000  
##                       Mean   :29.70   Mean   :0.5126   Mean   :0.4314  
##                       3rd Qu.:38.00   3rd Qu.:1.0000   3rd Qu.:1.0000  
##                       Max.   :80.00   Max.   :5.0000   Max.   :6.0000  
##  Survived
##  0:424
##  1:290
```

Figure 2. The summary table of the dataset. It appears that all the variables are on a similar scale.

The marginal effect of passenger's class on survival was checked via a bar plot. Passengers of the higher class (smaller in value) have obviously higher survival rate.

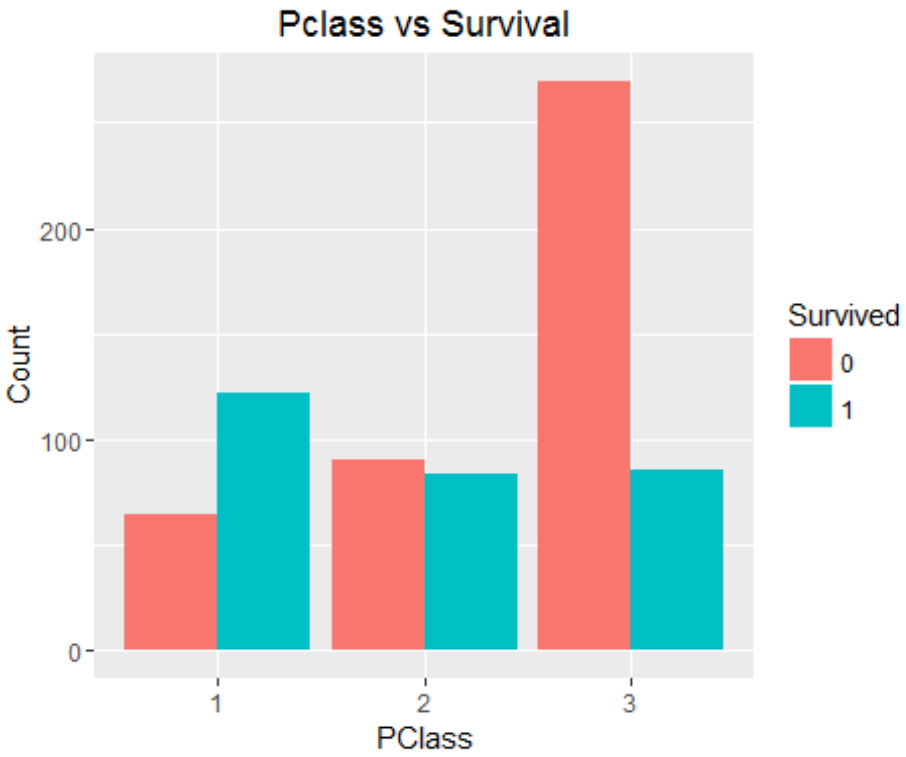

Figure 3. A bar plot showing the marginal effect of passenger class (Pclass) on survival.

The marginal effect of passenger's age on survival was checked via a box plot. The average age of passengers who survived the event is slightly lower than that of passengers who lost their lives in the shipwreck.

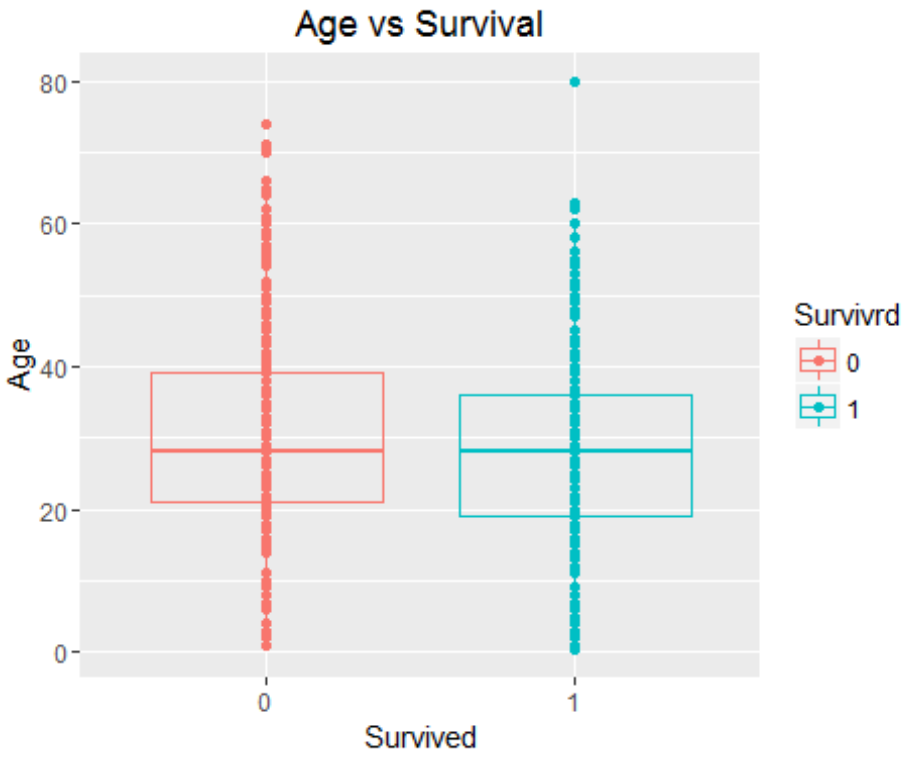

Figure 4. A box plot displaying the distribution of passenger age on passenger groups of both survived and not survived.

The marginal effect of passenger's gender on survival was checked via a contingency table with calculated survival rate. The survival rate of female is much higher than that of male, as expected.

```
gender_vs_survival
```

```
##          0   1 Survival_rate
## female  64 197     0.7547893
## male   360  93     0.2052980
```

Figure 5. A contingency table with calculated survival rate displaying the marginal effect of passenger's gender on survival.

The marginal effect of number of siblings and spouse on board on survival was checked via a contingency table with calculated survival rate. No clear pattern was recognized.

```
Sibsp_vs_survival
```

```
##     0   1 Survival_rate
## 0 296 175     0.3715499
## 1  86  97     0.5300546
## 2  14  11     0.4400000
## 3   8   4     0.3333333
## 4  15   3     0.1666667
## 5   5   0     0.0000000
```

Figure 6. A contingency table with calculated survival rate displaying the marginal effect of number of siblings or spouse on board on survival.

The marginal effect of number of parents and children on board on survival was checked via a contingency table with calculated survival rate. No clear pattern was recognized.

```
Parch_vs_survival
```

```
##     0   1 Survival_rate
## 0 335 186     0.3570058
## 1  49  61     0.5545455
## 2  29  39     0.5735294
## 3   2   3     0.6000000
## 4   4   0     0.0000000
## 5   4   1     0.2000000
## 6   1   0     0.0000000
```

Figure 7. A contingency table with calculated survival rate displaying the marginal effect of number of parents or children on board on survival.

2) Data split

The data set was arbitrarily split into two chunks: training and testing set with the ratio of 80% to 20%. After the split, the training set contains 571 observations and will be used to fit my models. Meanwhile the testing set contains 143 observations and I will test our models over it.

Models and results

1) Bayesian Model

a) Motivation

Bayesian models are designed to be appropriate to the data structure, without having to make approximation assumptions typical in Null Hypothesis Significance Testing (NHST). The inferences from a Bayesian analysis are richer and more informative than NHST because the posterior distribution reveals joint probabilities of combinations of parameter values. Moreover, there is no reliance on sampling distributions and p-values to interpret the parameter estimates.

b) Model, parameters and priors

All the categorical variables were replaced by their correspondent dummy variables before going into the model. Concretely, dummy variables Pclass.f2 and Pclass.f3 were created from Pclass and dummy variable Sex is created to replace the original categorical variable Sex.

The logit model is built via JAGS. The response Survived follows a Bernoulli distribution with the proportion of pi which is the odds. Odds pi is then linked with the linear predictor via the link function logit. beta0 is the parameter of intercept. beta1 to beta6 are the parameters for Pclass.f2 , Pclass.f3, Sex, Age, Sibsp and Parch respectively.

No prior knowledge of those parameters is available, so I set uniform priors of the same norm distribution with the mean equals 0 and the standard deviation equals 1.0E-6.

```
## DEFINE THE MODEL.
modelString = "
model{

# LOGIT MODEL
for (i in 1:n){
Survived[i] ~ dbern( pi[i] )
logit( pi[i] ) <- beta0 + beta1 * Pclass.f2[i] + beta2 * Pclass.f3[i] +
beta3 * Sex[i] + beta4 * Age[i] + beta5 * Sibsp[i] + beta6 * Parch[i]
}

# priors
beta0~dnorm( 0, 0.00001)
beta1~dnorm( 0, 0.00001)
beta2~dnorm( 0, 0.00001)
beta3~dnorm( 0, 0.00001)
beta4~dnorm( 0, 0.00001)
beta5~dnorm( 0, 0.00001)
beta6~dnorm( 0, 0.00001)
```

```
}
" # close quote for modelString
```

Figure 8. The model string used in JAGS.

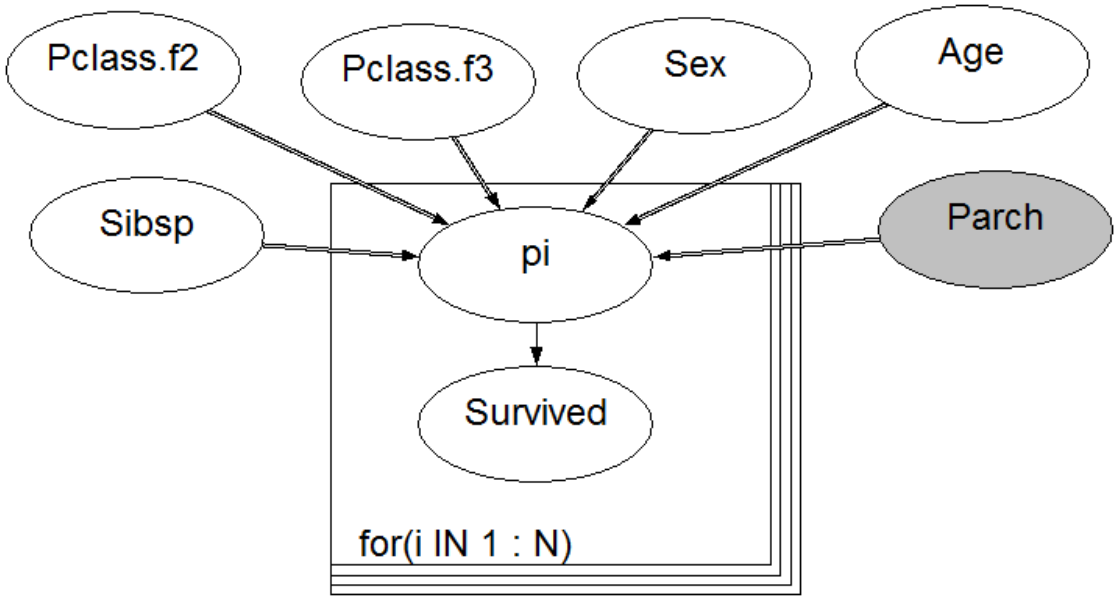

Figure 9. A doodle BUG that displays the structure of the model.

c) MCMC Diagnose

According to the trace plots, the chains overlap and mix with each other well. In addition, the Gelman-Rubin statistics, also known as shrinkage factors dropped rapidly and then remain at a low level close to 1 in our case. That also indicates good convergence. So it is believed that the MCMC samples I got were reliable representatives of the posterior distributions.

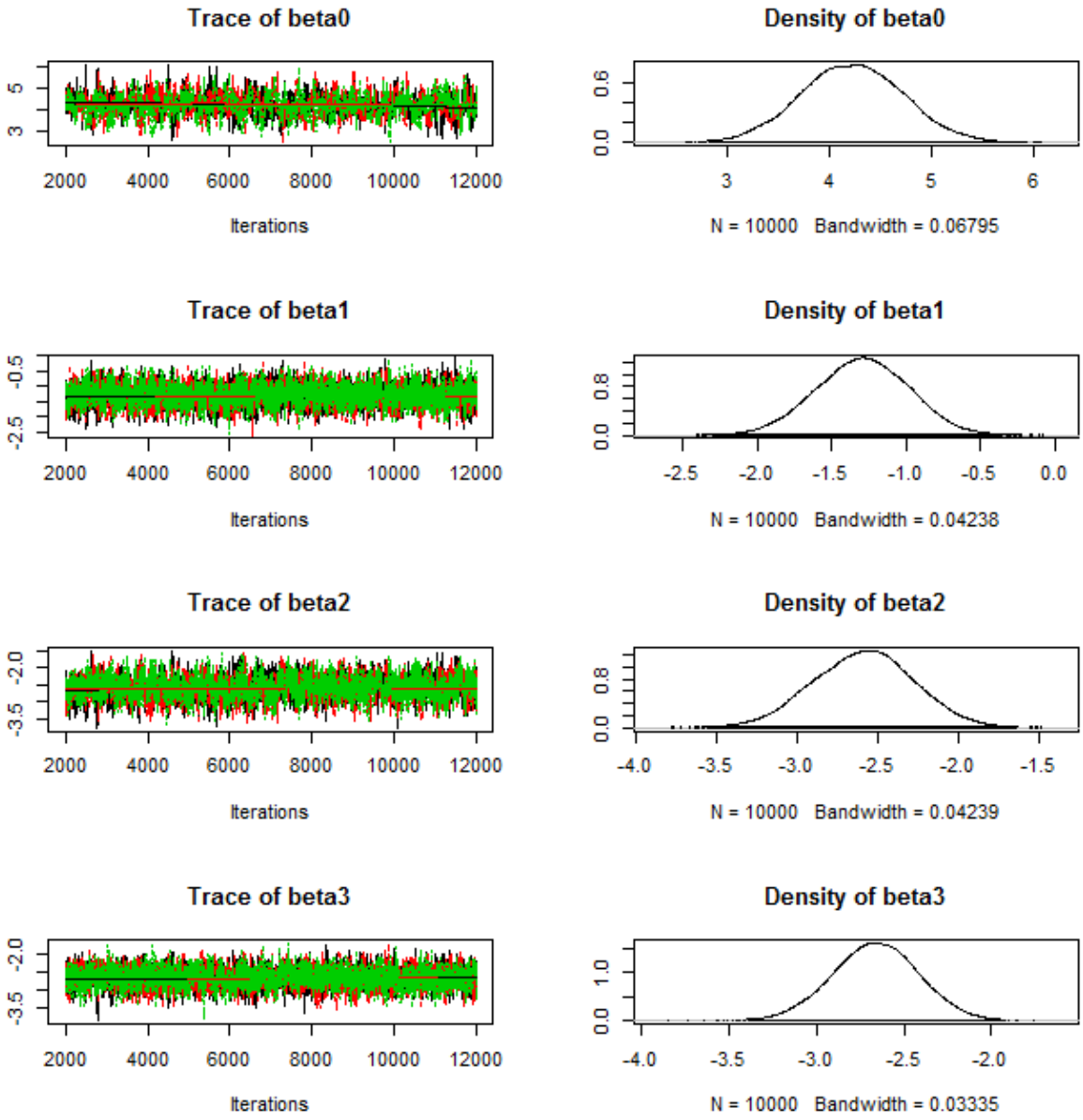

Figure 10. Trace plots and density plots of beta0 to beta3 from MCMC samples.

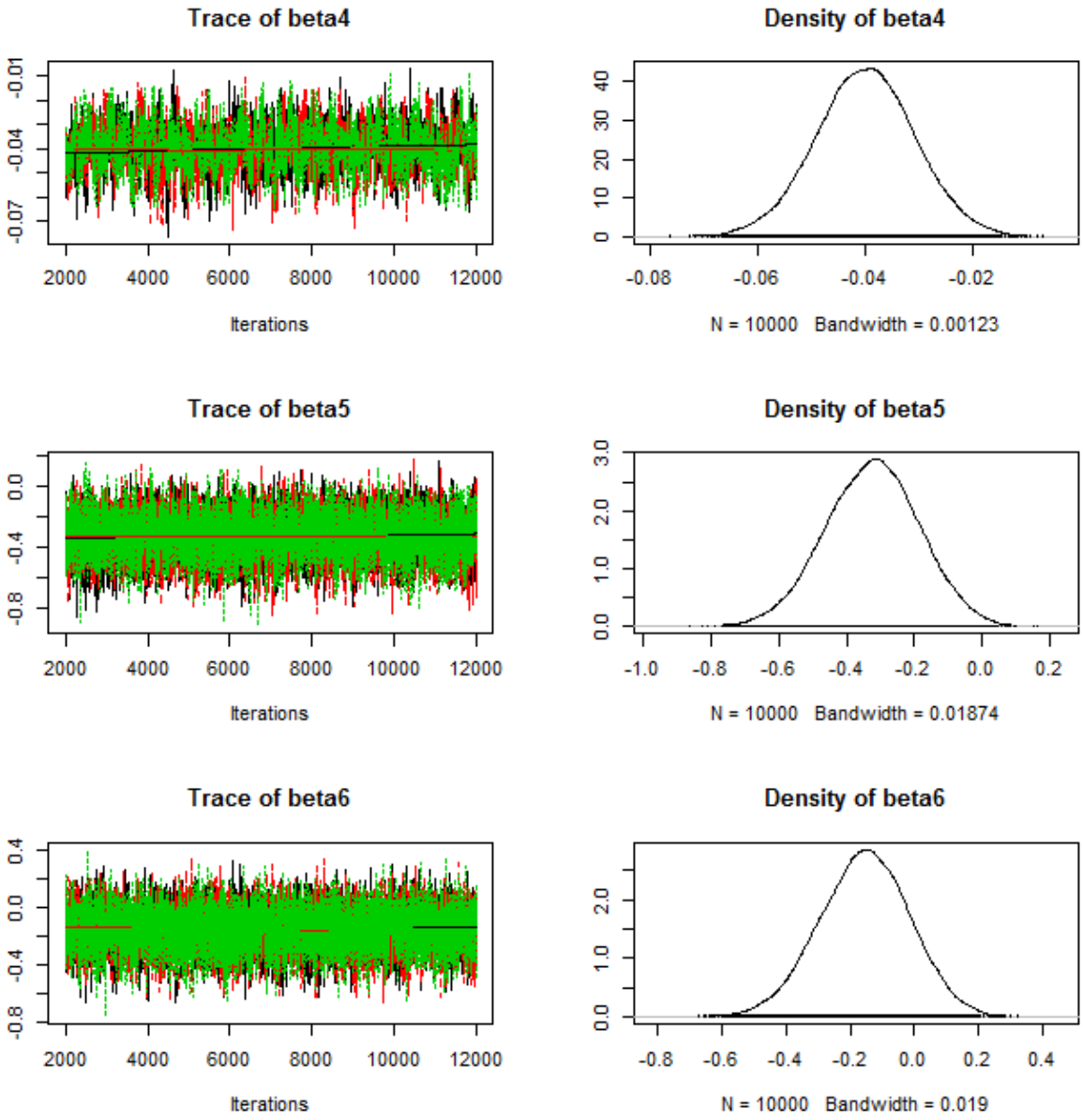

Figure 11. Trace plots and density plots of beta4 to beta6 from MCMC samples.

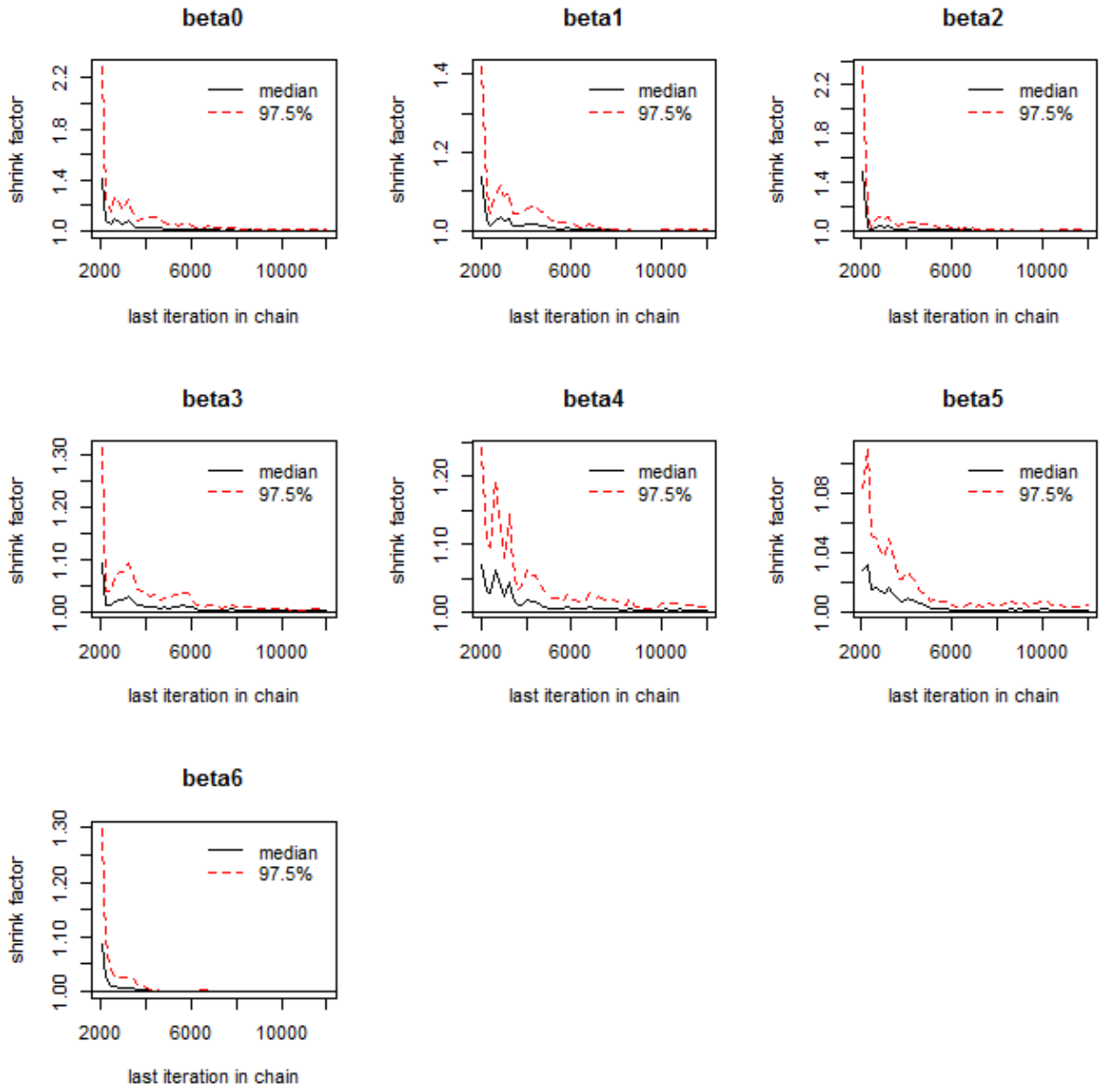

Figure 12. Gelman plots of beta0 to beta6 from MCMC samples.

d) Posterior estimates

The posterior estimates from MCMC is summarized below:

```
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
```

```
##           Mean       SD  Naive SE Time-series SE
## beta0  4.21140 0.488770 2.822e-03      0.0196832
## beta1 -1.29225 0.312292 1.803e-03      0.0069493
## beta2 -2.58790 0.309353 1.786e-03      0.0085102
## beta3 -2.65652 0.246900 1.425e-03      0.0046984
## beta4 -0.04007 0.008889 5.132e-05      0.0002852
## beta5 -0.32310 0.139010 8.026e-04      0.0019698

## beta6 -0.15672 0.141122 8.148e-04      0.0015287
##
## 2. Quantiles for each variable:
##
##           2.5%      25%      50%      75%    97.5%
## beta0  3.28009  3.87520  4.19625  4.53488  5.19509
## beta1 -1.90544 -1.50505 -1.29089 -1.08157 -0.68330
## beta2 -3.20098 -2.79476 -2.58174 -2.37862 -1.99329
## beta3 -3.14853 -2.82025 -2.65453 -2.48860 -2.17852
## beta4 -0.05787 -0.04597 -0.03982 -0.03406 -0.02304
## beta5 -0.60068 -0.41647 -0.32084 -0.22827 -0.05787
## beta6 -0.44222 -0.24850 -0.15391 -0.06060  0.10992
```

Figure 13. Summary of statistics from Bayesian logistic regression model.

Except for the beta0 for the intercept, all the other coefficients are negative. That means the model predicted negative effects of all the predictor (including dummy variables) variables on the odds of Survived. But the 95% HDI of beta6 is -0.44 to 0.11 including zero inside the range. So I believe that Parch might have less effect or even no effect on the odds of survival.

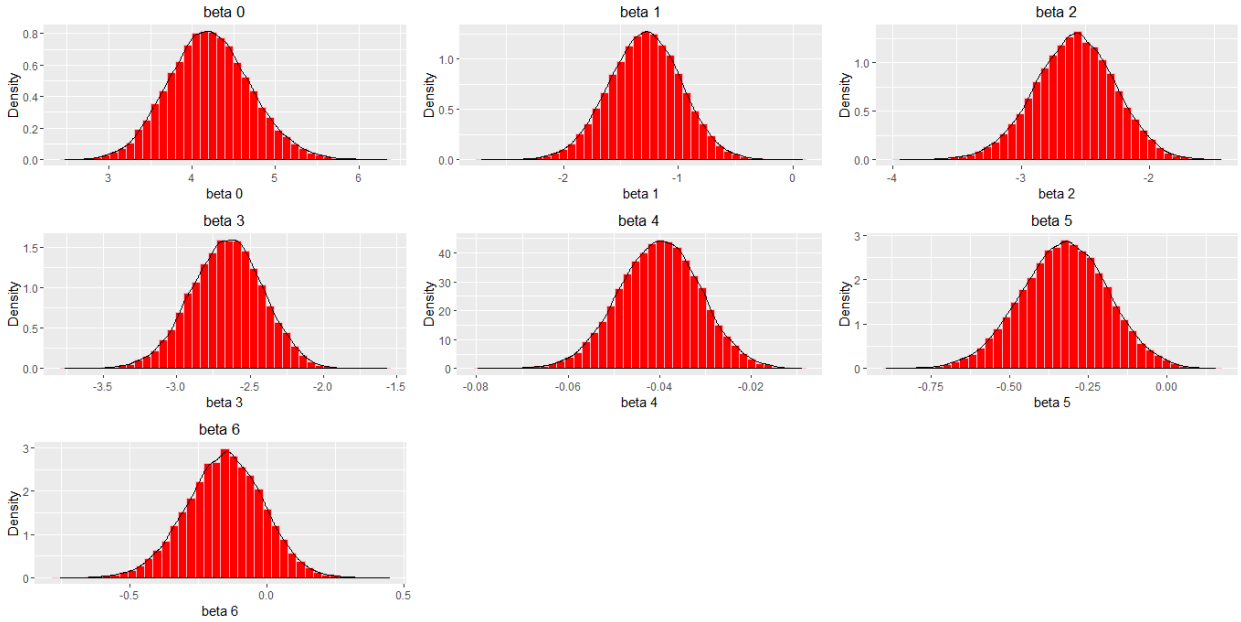

Figure 14. Density plots of beta0 to beta6 from MCMC samples.

```
pctg_change_odds
## [1] -0.72666376 -0.92520081 -0.93004697 -0.03928274 -0.27648889 -0.142311
```

Figure 15. Calculated percentage change of odds for each predictor variable.

From the point estimates, the percentage change of odds were calculated. It revealed that change from baseline to 1 in Pclass.f2 and Pclass.f3 would lead to 72.7% and 92.5% percent decrease respectively on odds. The change from baseline to 1 in Sex would lead to 93% percent decrease on odds. Then 1 unit increase in Age, Sibsp and Parch would lead to 3.9%, 27.6% and 14.2% percent decrease on odds. Basically, the model fitting suggested that passengers with higher class (smaller value) were more likely to survive. Female and younger people on board were more likely to survive. And passengers with less Siblings and spouse on board were more likely to survive. Since we discussed above there was not enough evidence to support the effect of Parch on response. The percentage change of odds on Parch was not convincing.

e) Model performance on testing set

The model was used to make prediction over the testing set. And the output was compared with the real response in the testing set. With a decision boundary of 0.5(If $P(y=1|X) > 0.5$ then $y = 1$ otherwise $y=0$.), the overall prediction accuracy from Bayesian model is 81.1%.

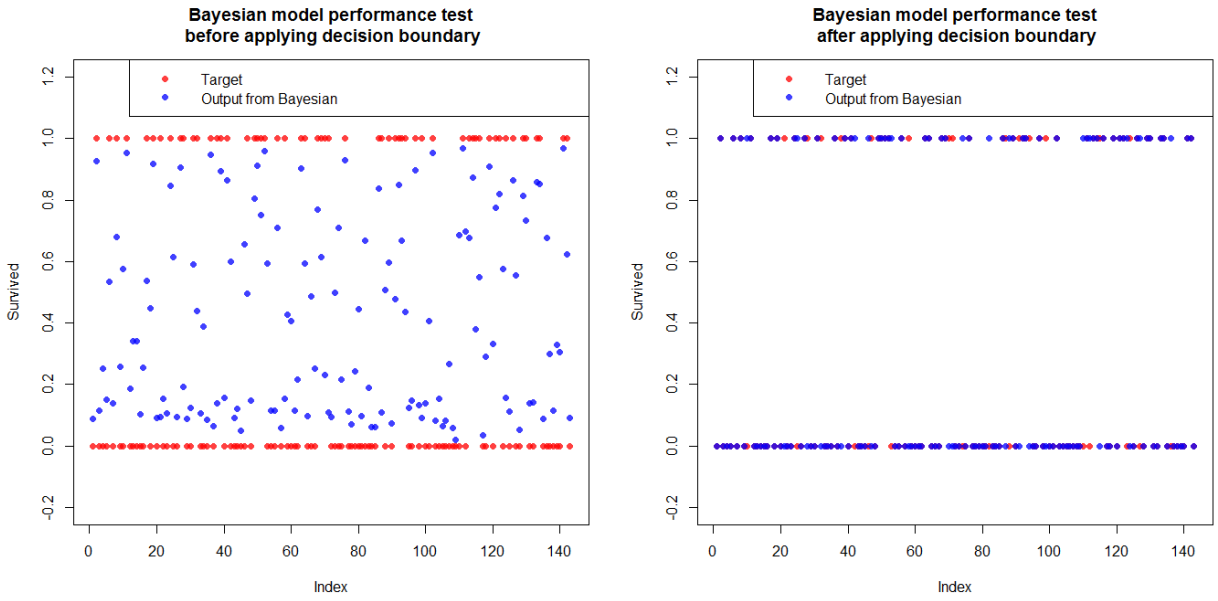

Figure 16. Bayesian model prediction performance on testing set.

2) Classical model

The classical model basically gave very similar result compared with Bayesian model. Only the effect of Parch on the survival is non-significant. And its prediction accuracy is 81.1%, exactly the same with the Bayesian model.

```
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = train[, c(2, 3, 5, 6, 7, 8)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6376  -0.6631  -0.3936   0.6413   2.3884
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.402561   0.597806   9.037  < 2e-16 ***
## Pclass      -1.272788   0.154460  -8.240  < 2e-16 ***
## Sexmale     -2.615407   0.244909 -10.679  < 2e-16 ***
## Age         -0.039146   0.008982  -4.358 1.31e-05 ***
## SibSp       -0.313939   0.136899  -2.293   0.0218 *
## Parch       -0.145229   0.137575  -1.056   0.2911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 772.89  on 570  degrees of freedom
## Residual deviance: 517.73  on 565  degrees of freedom
## AIC: 529.73
##
## Number of Fisher Scoring iterations: 5
```

Figure 17. Summary of statistics from classical logistic regression model.

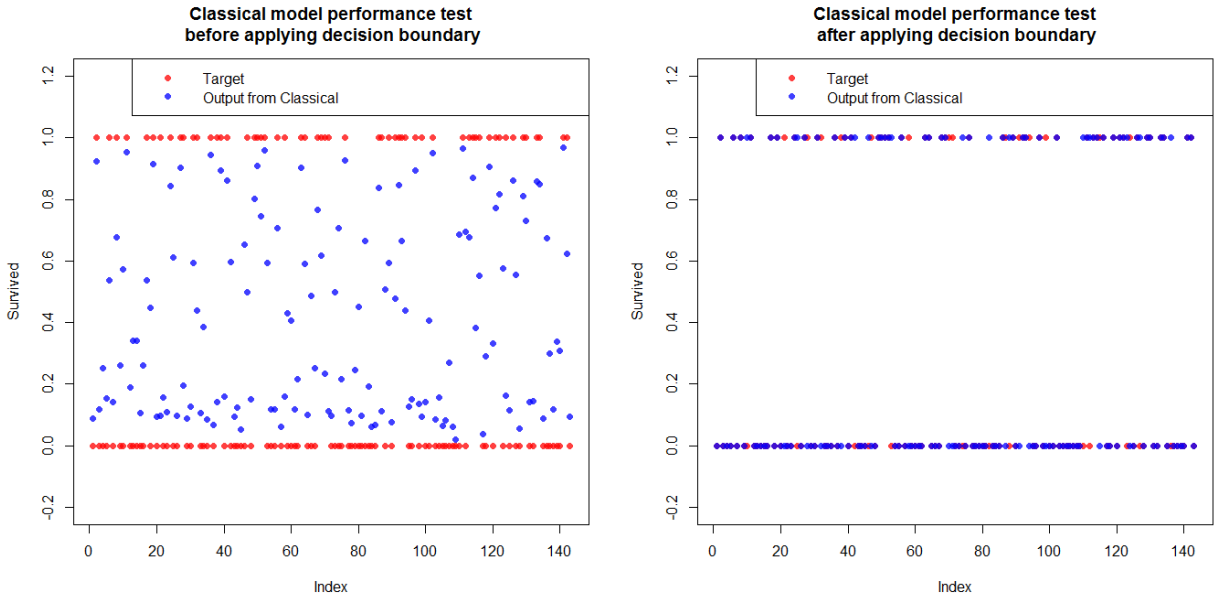

Figure 18. Classical model prediction performance on testing set.

3) Artificial Neural Network model

An artificial neural network (ANN) was established as well to learn the pattern from the training set. The ANN in my case used a default setting of 1 hidden layer and 10 hidden neurons on it. Its prediction performance on the testing set was examined to make comparison with Bayesian and classical models. Because of the randomness from initial weights generation, there were slight fluctuations on the trained weight if the model was run multiple times. The overall prediction accuracy is around 79% to 85%.

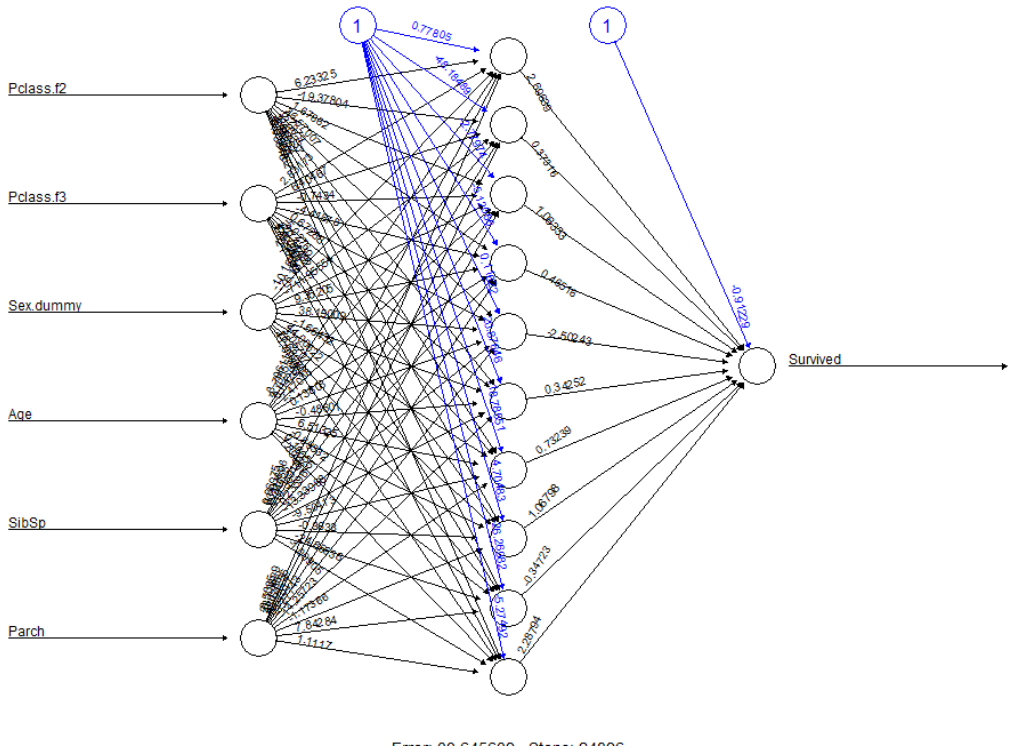

Figure 19. A schematic diagram of the trained ANN.

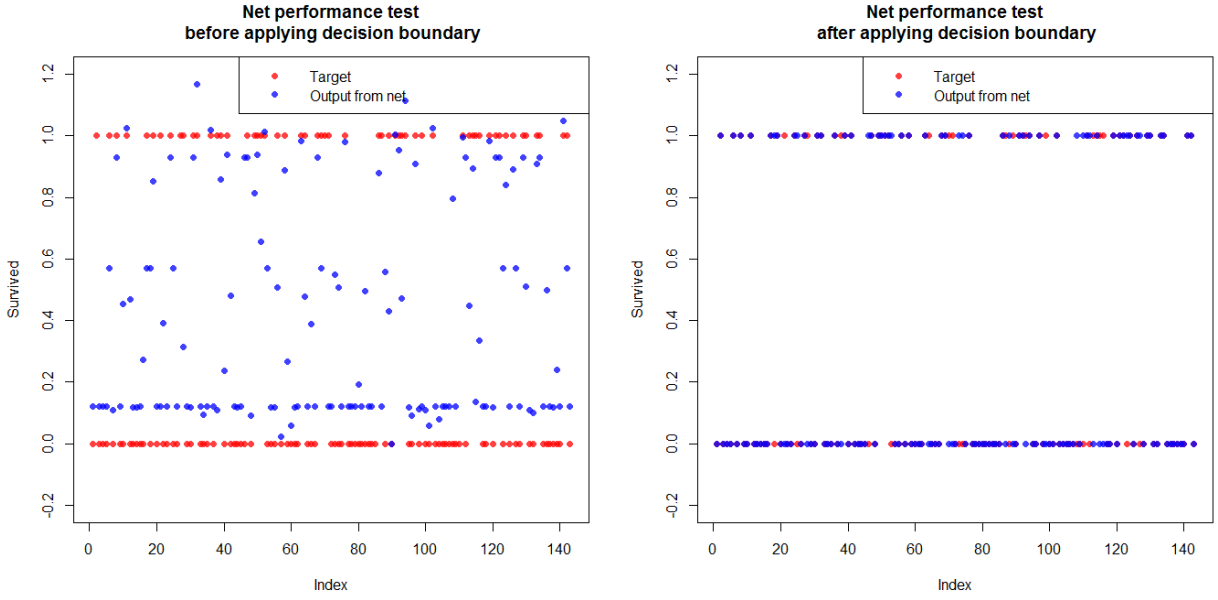

Figure 20. ANN model prediction performance on testing set.

Conclusion

Both Bayesian and classical models gave negative coefficients on (dummy variable) Sex and Age. My hypothesis was supported. So women and children were more likely to survive the shipwreck. In addition, the models also suggested that passengers with higher class (smaller value) and more siblings or spouse on board had a better chance to survive.

The prediction performance of Bayesian, classical models and ANN were all good and at a comparable level. However, these performance is somewhat dependent on the manual split of the data that I made earlier, therefore if we wish for a more precise score, we may need to run some kind of cross validation such as k-fold cross validation.