Bayesian Data Analysis Report on Titanic Dataset

Alp Gunsever

13/01/2021

1 - Introduction

This is a data analysis report intended to test bayesian data analysis skills using the titanic dataset from kaggle.

2- Exploratory Data Analysis

We are first going to make an exploratory analysis on the data available at hand. We format the data beforehand in both training and test sets.

The existing observations for different predictors will be formatted into data structures that can be handled by R in a meaningful way.

The following adjustments have been implemented to the training and test data sets together after they are combined into a single dataset:

- Passenger class predictor is transformed into factor type
- Titles have been extracted from the name variable and factored into Mr, Mrs, Miss, Master, Noble and Soldier levels.
- Cabin predictor is transformed into factor type.
- Embarked predictor is transformed into factor type.
- Ticket predictor is factored into L, F, P, C, A, S, W based on the letter they contain and into the number of digits if there is no letter in the ticket name. 3, 7, A, F, L, W types are combined into "Other" level as there are not so many observations in those factored levels.
- Family size predictor is created based on number of siblings and spouses including self. Family type predictor is created based on the family size predictor. If the number of family members are equal to 1, then family type is assigned to "Singleton". If family size is between 1 and 4, then family type is assigned to "small". If family size is greater than 4, then family type is assigned to "large". Family type is factored into these 3 levels.
- Sex predictor is changed to isMale and factored as a binary predictor.

The data summary can be seen below:

```
##   PassengerId  Pclass        Age          Ticket         Fare
##  Min.   :   1   1:323   Min.   : 0.17   4    :144   Min.   :  0.000
##  1st Qu.: 328   2:277   1st Qu.:21.00   5    :193   1st Qu.:  7.896
##  Median : 655   3:709   Median :28.00   6    :596   Median : 14.454
##  Mean   : 655           Mean   :29.88   C    : 88   Mean   : 33.295
##  3rd Qu.: 982           3rd Qu.:39.00   Other: 99   3rd Qu.: 31.275
##  Max.   :1309           Max.   :80.00   P    :132   Max.   :512.329
##                         NA's   :263     S    : 57   NA's   :1
##      Cabin      Embarked       title           fSize     isMale
##  C      :  94   C   :270   Master : 61   large    : 82   0:466
##  B      :  65   Q   :123   Miss   :266   singleton:790   1:843
##  D      :  46   S   :914   Mr     :773   small    :437
##  E      :  41   NA's:  2   Mrs    :197
##  A      :  22              Noble  :  5
##  (Other):  27              Soldier:  7
##  NA's   :1014
```

```
## 'data.frame':    1309 obs. of  10 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ Ticket     : Factor w/ 7 levels "4","5","6","C",..: 5 6 7 3 3 3 2 3 3 3 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 8 levels "A","B","C","D",..: NA 3 NA 3 NA NA 5 NA NA NA ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...
##  $ title      : Factor w/ 6 levels "Master","Miss",..: 3 4 2 4 3 3 3 1 4 4 ...
##  $ fSize      : Factor w/ 3 levels "large","singleton",..: 3 3 2 3 2 2 2 1 3 3 ...
##  $ isMale     : Factor w/ 2 levels "0","1": 2 1 1 1 2 2 2 2 1 1 ...
```

Cabin and age predictors have a lot of missing observations so it can be beneficial to get rid of these predictors but we will also be losing a lot of information so it makes sense to further investigate if imputation is possible.

We will impute the missing age observations by replacing them with median of each title group that the individual belongs to.

Individuals that are upper class passengers, are members of small families(2 to 4 family members), embarked from C and are below 60 years old tend to have higher survival rates.

Cabin variable will be dropped as it doesn't seem possible to impute this predictor with lots of missing observations. However, it might still makes sense to impute age predictor.

The imputed dataset will be used for analysis.

3- Prior Predictive Checking

We will perform a prior predictive check with only using the priors and no data.The reason for making prior predictive analysis is to make a sanity check on the priors without using the likelihood.

We have used weakly informative robust prior of student_t(3,0,2.5) for both population-level and group-level parameters. We also used lkj(2) for the correlation matrix. It can be seen that the spread for prior predictive samples have much higher variance than the actual response variable for the training set. It confirms that the chosen prior can be used for posterior predictive check along with the likelihood.

4 - Model Fitting and Algorithm Diagnostics

5 - Posterior Predictive Checking

6 - Additional Models and Model Improvements

7 - Model Comparison

8 - Prediction Submission